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 elucidate the origins of cellular diversity and assist the annotation of new cell atlases. Yet, comparison between distant relatives is hindered by complex gene histories and diversifications in expression programs. Previously, we introduced the self-assembling manifold (SAM) algorithm to robustly reconstruct manifolds from single-cell data (Tarashansky et al., 2019). Here, we build on SAM to map cell atlas manifolds across species. This new method, SAMap, identifies homologous cell types with shared expression programs across distant species within phyla, even in complex examples where homologous tissues emerge from distinct germ layers. SAMap also finds many genes with more similar expression to their paralogs than their orthologs, suggesting paralog substitution may be more common in evolution than previously appreciated. Lastly, comparing species across animal phyla, spanning mouse to sponge, reveals ancient contractile and stem cell families, which may have arisen early in animal evolution.

Introduction

There is much ongoing success in producing single-cell transcriptomic atlases to investigate the cell type diversity within individual organisms (Regev et al., 2017). With the growing diversity of cell atlases across the tree of life (Briggs et al., 2018; Cao et al., 2019; Fincher et al., 2018; Hu et al., 2020; Musser et al., 2019; Plass et al., 2018; Siebert et al., 2019; Wagner et al., 2018), a new frontier is emerging: the use of cross-species cell type comparisons to unravel the origins of cellular diversity and uncover species-specific cellular innovations (Arendt et al., 2019; Shafer, 2019). Further, these comparisons promise to accelerate cell type annotation and discovery by transferring knowledge from well-studied model organisms to under-characterized animals.

However, recent comparative single-cell analyses are mostly limited to species within the same phylum (Baron et al., 2016; Geirsdottir et al., 2019; Sebé-Pedrós et al., 2018; Tosches et al., 2018). Comparisons across longer evolutionary distances and across phyla are challenging for two major reasons. First, gene regulatory programs diversify during evolution, diminishing the similarities in cell type specific gene expression patterns. Second, complex gene evolutionary history causes distantly related organisms to share few one-to-one gene orthologs (Nehrt et al., 2011), which are often relied upon for comparative studies (Briggs et al., 2018; Shafer, 2019). This effect is compounded by the growing evidence suggesting that paralogs may be more functionally similar than orthologs across species, due to differential gain (neo-functionalization), loss (non-functionalization), or partitioning (sub-functionalization) events among paralogs (Nehrt et al., 2011; Prince & Pickett, 2002; Stamboulian et al., 2020; Studer & Robinson-Rechavi, 2009).

Here, we present the Self-Assembling Manifold mapping (SAMap) algorithm to enable mapping single-cell transcriptomes between phylogenetically remote species. SAMap relaxes the constraints imposed by sequence orthology, using expression similarity between mapped cells to infer the relative contributions of homologous genes, which in turn refines the cell type mapping. In addition, SAMap uses a graph-based data integration technique to identify reciprocally connected cell types across species with greater robustness than previous single-cell data integration methods (Haghverdi et al., 2018; Hie et al., 2019; Polański et al., 2019; Stuart et al., 2019).

Using SAMap, 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**). 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 densely 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 often exhibit differential expression of orthologs and similar expression of paralogs, suggesting that the substitution and swapping of paralogs in cell types may be more common in evolution than previously appreciated. Overall, our study represents an important step towards analyzing the evolutionary origins of specialized cell types and their associated gene expression programs in animals.

Results

The SAMap algorithm

SAMap iterates between two modules. The first module constructs a gene-gene bipartite graph with cross-species edges connecting homologous gene pairs, initially weighted by protein sequence similarity (**Figure 1C**). In the second module, SAMap uses the gene-gene graph to project the two single-cell transcriptomic datasets into a joint, lower-dimensional manifold representation, from which each cell's mutual cross-species neighbors are linked to stitch the cell atlases together (**Figure 1D**). Then, using the joint manifold, the expression correlations between homologous genes are computed and used to reweight the edges in the gene-gene homology graph in order to relax SAMap's initial dependence on sequence similarity. The new homology graph is used as input to the subsequent iteration of SAMap, and the algorithm continues until convergence, defined as when the cross-species mapping does not significantly change between iterations (**Figure 1E**).

This algorithm overcomes several challenges inherent to mapping single-cell transcriptomes between distantly related species. First, complex gene evolutionary history often results in many-to-many homologies with convoluted functional relationships (Briggs et al., 2018; Nehrt et al., 2011). SAMap accounts for this by using the full homology graph to project each dataset into both its own and its partner's respective principal component (PC) spaces, constructed by the SAM algorithm, which we previously developed to robustly and sensitively identify cell types (Tarashansky et al., 2019). The resulting within- and cross-species projections are concatenated to form the joint space.

For the 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. Iteratively refining the homology graph to only include positively correlated gene pairs prunes the many-to-many homologies to only include genes that are expressed in the same mapped cell types.

Second, frequent gene losses and the acquisitions of new genes result in many cell type gene expression signatures being species-specific, limiting the amount of information that is comparable across species. Restricting the analysis of each dataset to only include genes that are shared across species would result in a decreased ability to resolve cell types and subtypes with many species-specific gene signatures. SAMap solves this problem by constructing the joint space through the concatenation of within- and cross-species projections, thus encoding all genes from both species.

Lastly, the evolution of expression programs gradually diminishes the similarity between homologous cell types. To account for this effect, SAMap links cell types across species while tolerating their differences. Cells are mapped by calculating each of their *k* mutual nearest cross-species neighbors in the combined projection. To establish more robust mutual connectivity, we integrate information from each cell's local, within-species neighborhood (**Figure 1D**), overcoming the inherent stochasticity of cross-species correlations. Two cells are thus defined as mutual nearest cross-species neighbors when their respective neighborhoods have mutual connectivity. It is important to note that the magnitude of connections is not directly calculated from their expression similarity, allowing cell types with diverged expression profiles to be tightly linked if they are among each other's closest cross-species neighbors.

Paralog substitutions are prevalent between homologous cell types in frog and fish

We first applied SAMap to the *Xenopus* and zebrafish atlases, which both encompass embryogenesis until early organogenesis (Briggs et al., 2018; Wagner et al., 2018). Previous analysis had linked cell types between these two organisms by matching ontogeny, thereby providing a reference for comparison. 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**). We measured the mapping strength between cell types by calculating an alignment score (edge width 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.

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**) (Briggs et al., 2018). 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 matched time points along several cell lineages (**Figure 2C**). While the concordance was consistent across cell types, we noticed that the exact progression of developmental timing can vary, suggesting that SAMap can quantify heterochrony with cell type resolution.

SAMap also weakly linked several closely related cell types with different ontogeny. For example, optic cells from both species are also connected to eye primordium, frog skeletal muscles to fish presomitic mesoderm, and frog hindbrain to fish forebrain/midbrain. Notable exceptions also included mapped 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* (Dubaissi et al., 2014), *grhl* (Miles et al., 2017)) and proteins involved in vesicular protein trafficking (**Figure 2 – figure supplement 1**). This observation supports the notion that cell types may be transcriptionally and evolutionarily related despite having different developmental origins (Arendt et al., 2016).

To benchmark SAMap performance, we used Eggnog (Huerta-Cepas et al., 2019) to define one-to-one vertebrate orthologs between fish and frog and fed these gene pairs as input to several broadly used single-cell data integration methods, Seurat (Stuart et al., 2019), Liger (Welch et al., 2019), Harmony (Korsunsky et al., 2019), Scanorama (Hie et al., 2019), and BBKNN (Polański et al., 2019). We found that they failed to map the two atlases, yielding minimal alignment between them (**Figure 2D** and **Figure 2 – figure supplement 2**). We also compared the results when restricting SAMap to using the one-to-one orthologs instead of the full homology graph. Even when removing the many-to-many gene homologies and the iterative refinement of the homology graph, we identified similar, albeit weaker, cell type mappings. This suggests that, at least for the frog and fish comparison, SAMap's performance is owed in large part to its robust, atlas stitching approach.

The key benefit of using the full homology graph is to enable the systematic identification of gene paralogs that exhibit greater similarity in expression across species than their corresponding orthologs. These events are expected to arise as the result of gene duplications followed by diversification of the resulting in-paralogs (Studer & Robinson-Rechavi, 2009). In addition, genetic compensation by transcriptional adaptation, where loss-of-function mutations are balanced by upregulation of related genes with similar sequences (El-Brolosy et al., 2019), could also result in this signature. In total, SAMap selected 8,286 vertebrate orthologs and 7,093 eukaryotic paralogs, as enumerated by Eggnog, for manifold alignment. Among these, 565 genes have markedly higher expression correlations (correlation difference > 0.3) with their paralogs than their orthologs (**Figure 2E** and **Figure 2 – figure supplement 3**), and 209 genes have orthologs that are either completely absent or lowly-expressed with no cell-type specificity (**Supplementary Table 3**), suggesting that they may have lost their functional roles at some point and were compensated for by their paralogs. We term these events as "paralog substitutions". SAMap linked an additional 297 homologous pairs not previously annotated by orthology or paralogy, but which exhibited sequence similarity and high expression correlations (>0.5 Pearson correlation). These likely represent unannotated orthologs/paralogs or isofunctional, distantly related homologs (Gabaldón & Koonin, 2013). These results illustrate the potential of SAMap in leveraging single-cell gene expression data for pruning the networks of homologous genes to identify evolutionary substitution of paralogs and, more generally, identify non-orthologous gene pairs that may perform similar functions in the cell types within which they are expressed.

Homologous cell types between two flatworm species with divergent body plans

To test if we can identify homologous cell types in animals with radically different body plans, we mapped the cell atlases of two flatworms, the planarian *Schmidtea mediterranea* (Fincher et al., 2018), and the trematode *Schistosoma mansoni*, which we collected recently (Li et al., 2020). They represent two distant lineages within the same phylum but have remarkably distinct body plans and autecology (Laumer et al., 2015; Littlewood & Waeschenbach, 2015). While planarians live in freshwater and are known for their ability to regenerate (Reddien, 2018), schistosomes live as parasites in humans. The degree to which cell types are conserved between them is unresolved, given the vast phenotypic differences caused by the transition from free-living to parasitic habits (Laumer et al., 2015).

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 (Fincher et al., 2018) (**Figure 3A**). These mappings are supported by both known cell type specific marker genes and numerous homologous transcriptional regulators (**Figure 3B** and **Figure 3 – figure supplement 1**).

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 (Fincher et al., 2018; Zeng et al., 2018). By mapping the schistosome neoblasts to a planarian neoblast atlas (Zeng et al., 2018), we found that the schistosome has a population of neoblasts (ε-cells (Wang et al., 2018)) that cluster with the planarian's pluripotent neoblasts, both expressing a common set of TFs (e.g., *soxp2, unc4*, *pax6a*, *gcm1*) (**Figure 3C-D**). The ε-cells are closely associated with juvenile development and lost in adult schistosomes (Wang et al., 2018), indicating pluripotent stem cells may be a transient population restricted to their early developmental stages. This is consistent with the fact that, whereas schistosomes can heal wounds, they have limited regenerative ability (Wendt & Collins, 2016). SAMap also linked other schistosome neoblast populations with planarian progenitors, including two populations of schistosome neoblasts (denoted as μ (Tarashansky et al., 2019) and μ') to planarian muscle progenitors, all of which express *myoD*, a canonical master regulator of myogenesis (Scimone et al., 2017). These likely represent early and late muscle progenitors, respectively, as μ-cells do not yet express differentiated muscle markers such as *troponin*, whereas μ'-cells do (**Figure 3 – figure supplement 2).**

Cell type families spanning the animal tree of life

To compare cell types across broader taxonomic scales, we extended our analysis to include juvenile freshwater sponge (*Spongilla lacustris*) (Musser et al., 2019), adult *Hydra* (*Hydra vulgaris*) (Siebert et al., 2019), and mouse (*Mus musculus*) embryogenesis (Pijuan-Sala et al., 2019) atlases. In total, SAMap linked 1,051 cross-species pairs of cell types, defined by the annotations used in each respective study. 95% of the cell type pairs are supported by at least 40 enriched gene pairs, and 87% are supported by more than 100 gene pairs, indicating that SAMap does not spuriously connect cell types with limited overlap in expression profiles (**Figure 4A**).

We next extended the notion of cell type pairs to cell type trios, as mapped cell types gain additional support if they share transitive relationships to other cell types through independent mappings, forming cell type triangles among species. The transitivity of a cell type pair (edge) or a cell type (node) can be quantified as the fraction of triads to which they belong that form triangles (**Figure 4B**). The majority (81%) of cell type pairs have non-zero transitivity independent of alignment score and the number of enriched gene pairs (**Figure 4 – figure supplement 1-2**). Cell type pairs with fewer than 40 enriched gene pairs tend to have lower (<0.4) transitivity (**Figure 4 – figure supplement 2**). In addition, 16% of mapped cell type pairs have zero edge transitivity but non-zero node transitivity, meaning that at least one of the cell types connects to only a single member of an interconnected cell type group (**Figure 4C**). Such edges may be of lower confidence as they should connect to other members of the same group and are thus excluded from downstream analysis.

Among the interconnected groups of cell types, we identified families of contractile cells and neural cells (**Figure 4D**). Both cell type families are highly transitive compared to the overall graph transitivity (bootstrap p-value < $1x10^{-5}$), meaning that their constituent cell types have more transitive edges within the group than outside the group (**Figure 4E**). In addition, the dense, many-to-many connections within the contractile and neural families are each supported by at least 40 enriched gene pairs (**Figure 4F**). Consistent with the nerve net hypothesis suggesting a unified origin of neural cell types (Tosches & Arendt, 2013), the neural family includes vertebrate brain tissues, both bilaterian and cnidarian neurons, cnidarian nematocytes that share the excitatory characteristics of neurons (Weir et al., 2020), and *Spongilla* choanocytes and apopylar cells, both of which are not considered as neurons but have been shown to express postsynaptic-like scaffolding machinery (Musser et al., 2019; Wong et al., 2019). The contractile family includes myocytes in bilaterian animals, *Hydra* myoepithelial cells that are known to have contractile myofibrils (Buzgariu et al., 2015), and sponge pinacocytes and myopeptidocytes, both of which have been implicated to play roles in contractility (Musser et al., 2019; Sebé-Pedrós et al., 2018). In contrast to the families encompassing all seven species, we also found a fully interconnected group that contains invertebrate pluripotent stem cells, including planarian and schistosome neoblasts, *Hydra* interstitial cells, and sponge archeocytes (Alié et al., 2015). The lack of one-to-one connections across phyla is in keeping with recent hypotheses that ancestral cell types diversified into families of cell types after speciation events (Arendt et al., 2016, 2019). Our findings thus suggest that these cell type families diversified early in animal evolution.

Transcriptomic signatures of cell type families

The high interconnectedness between cell types across broad taxonomic scales suggests that they should share ancestral transcriptional programs (Arendt et al., 2016; Tosches et al., 2018). SAMap identified broad transcriptomic similarity between bilaterian and nonbilaterian contractile cells that extends beyond the core contractile apparatus. It links a total of 23,601 gene pairs, connecting 5,471 unique genes, which are enriched in at least one contractile cell type pair. Performing functional enrichment analysis on these genes, we found cytoskeleton and signal transduction functions to be enriched ($p\text{-value} < 10^{-3}$) based on the KOG functional classifications (Tatusov et al., 2003) 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 (Brunet et al., 2016) 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 (Larroux et al., 2008) enriched in contractile cells from all seven species. The Fox proteins included FoxC, which is known to regulate cardiac muscle identity in vertebrates (Brunet et al., 2016) 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 (**Figure 5 – figure supplement 1**), suggesting that FoxG may play an underappreciated role in contractile cell specification outside vertebrates.

For the family of invertebrate stem cells, we identified 3,343 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 via evolutionary splitting that results in separate derived programs. While this notion of coupled cellular and molecular evolution has gained significant traction in the past years, systematically comparing cell type-specific gene expression programs across species has remained a challenging problem. Here, we map single-cell atlases between evolutionarily distant species in a manner that accounts for the complexity of gene evolution. 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 findings support the notion that cell types evolve via hierarchical diversification (Arendt et al., 2019), resulting in cell type families composed of evolutionarily related cell types sharing a regulatory gene expression program that originated in their common ancestor. One-to-one cell type homologies should exist only if no further cell type diversification has occurred since the speciation.

In parallel, SAMap systematically identifies instances where paralogs exhibit greater expression similarity than orthologs across species. Paralog substitution likely occurs due to differential loss or retention of cell type-specific expression patterns of genes that were duplicated in the common ancestor (Studer & Robinson-Rechavi, 2009). Alternatively, paralog substitutions could arise due to compensating upregulation of paralogs following a loss-of-function mutation acquired by an ortholog (El-Brolosy et al., 2019). While the analysis presented here focuses on comparisons between two species, incorporating multiple species into a single analysis that also accounts for their phylogenetic relatedness could enable determining the specific order of paralog substitutions, associated cell type diversification events, and the mechanism by which they arose. However, this would require cell atlases that consistently sample key branching points along the tree of life. Nevertheless, identifying lineage-specific paralog substitution signatures should be accessible in extensively studied vertebrate single-cell atlases, as the vertebrate clade is where existing data and knowledge are most concentrated.

Besides applications in evolutionary biology, we anticipate SAMap can catalyze the annotation of new cell atlases from non-model organisms, which often represents a substantial bottleneck requiring extensive manual curation and prior knowledge. 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. Moreover, our approach allows leveraging existing and forthcoming single-cell gene expression data to predict functionally similar gene homologs, which can serve as guideposts for mechanistic molecular studies.

Materials and Methods

Data and Code Availability

- The source code for SAMap is publicly available at Github (https://github.com/atarashansky/SAMap), along with the code to perform the analysis and generate the figures.
- The datasets analyzed in this study are detailed in **Supplementary Table 1** with their accessions, and annotations provided.

The SAMap Algorithm

The SAMap algorithm contains three major steps: preprocessing, mutual nearest neighborhood alignment, and gene-gene correlation initialization. The latter two are repeated for three iterations, by default, to balance alignment performance and computational runtime. SAMap runs up to one hour on an average desktop computer for 200,000 total cells.

1. Preprocessing.

1.1. Generate gene homology graph via reciprocal BLAST.

We first construct a gene-gene bipartite graph between two species by performing reciprocal BLAST of their respective transcriptomes using *tblastx*, or proteomes using *blastp*. *tblastn* and *blastx* are used for BLAST between proteome and transcriptome. When a pair of genes share multiple High Scoring Pairs (HSPs), which are local regions of matching sequences, we use the HSP with the highest bit score to measure homology. Only pairs with E-value < 10^{-6} are included in the graph.

Here we define similarity using BLAST, though SAMap is compatible with other protein homology detection methods (e.g. HMMER (Eddy, 2008)) or orthology inference tools (e.g. OrthoClust (Yan et al., 2014) and Eggnog (Huerta-Cepas et al., 2019)). While each of these methods has known strengths and limitations, BLAST is chosen for its broad usage, technical convenience, and compatibility with low-quality transcriptomes.

We encode the BLAST results into two triangular adjacency matrices, A and B, each containing bit scores in one BLAST direction. We combine A and B to form a gene-gene adjacency matrix G. After symmetrizing G, we remove edges that only appear in one direction: $G = Recip(\frac{1}{2}[(A+B)+(A+B)^T]) \in \Re^{m_1+m_2 \times m_1+m_2}$, where $Recip$ only keeps reciprocal edges, and m_1 and m_2 are the number of genes of the two species, respectively. To filter out relatively weak homologies, we also remove edges where $G_{ab} < 0.25 \underset{b}{max}(G_{ab})$. Edge weights are then normalized by the maximum edge weight for each gene and transformed by a hyperbolic tangent function to increase discriminatory power between low and high edge weights, $\hat{G}_{ab} = 0.5 + 0.5\, tanh\,(10G_{ab}/\underset{b}{max}(G_{ab}) - 5)$.

1.2. Construct manifolds for each cell atlas separately using the SAM algorithm.

The scRNAseq datasets are normalized such that each cell has a total number of raw counts equal to the median size of single-cell libraries. Gene expressions are then log-normalized with the addition of a pseudocount of 1. Genes expressed (i.e.,$log_2(D+1) >$

1) in greater than 96% of cells are filtered out. SAM is run using the following parameters: *preprocessing* = *'StandardScaler'*, *weight_PCs* = *False*, *k* = *20*, and *npcs* = *150*. A detailed description of parameters is provided previously (Tarashansky et al., 2019). SAM outputs N_1 and N_2, which are directed adjacency matrices that encode *k*-nearest neighbor graphs for the two datasets, respectively.

SAM only includes the top 3,000 genes ranked by SAM weights and the first 150 principal components (PCs) in the default mode to reduce computational complexity. However, downstream mapping requires PC loadings for all genes. Thus, in the final iteration of SAM, we run PCA on all genes and take the top 300 PCs. This step generates a loading matrix for each species i, $L_i \in \mathfrak{R}^{300\times m_i}$.

2. Mutual nearest neighborhood alignment.

2.1. Transform feature spaces between species.

For the gene expression matrices $Z_i \in \mathfrak{R}^{n_i\times m_i}$, where n and m are the number of cells and genes respectively, we first zero the expression of genes that do not have an edge in $\hat{G}$ and standardize the expression matrices such that each gene has zero mean and unit variance, yielding $\tilde{Z}_i$. $\hat{G}$ represents a bipartite graph in the form of $\hat{G} = \begin{bmatrix} 0_{m_1,m_1} & H \in \mathfrak{R}^{m_1\times m_2} \\ H^T \in \mathfrak{R}^{m_2\times m_1} & 0_{m_2,m_2} \end{bmatrix}$, where $0_{m,m}$ is $m \times m$ zero matrix and H is the biadjacency matrix. Letting $H_1 = H$ and $H_2 = H^T$ encoding directed edges from species 1 to 2 and 2 to 1, respectively, we normalize the biadjacency matrix H_i such that each row sums to 1: $\hat{H}_i = SumNorm(H_i) \in \mathfrak{R}^{m_i\times m_j}$, where the $SumNorm$ function normalizes the rows to sum to 1. The feature spaces can be transformed between the two species via weighted averaging of gene expression, $\tilde{Z}_{ij} = \tilde{Z}_i \hat{H}_i$.

2.2. Project single-cell gene expressions into a joint PC space.

We project the expression data from two species into a joint PC space (Barkas et al., 2019), $P_i = \tilde{Z}_i {L_i}^T$ and $P_{ij} = \tilde{Z}_{ij} {L_j}^T$. We then horizontally concatenate the principal components P_i and P_{ij} to form $\hat{P}_i \in \Re^{n_i \times 600}$.

2.3. Calculate k-nearest cross-species neighbors for all cells.

Using the joint PCs, $\hat{P}_i$, we identify for each cell the k-nearest neighbors in the other dataset using cosine similarity ($k = 20$ by default). Neighbors are identified using the *hnswlib* library, a fast approximate nearest-neighbor search algorithm (Malkov & Yashunin, 2020). This outputs two directed biadjacency matrices $C_i \in \Re^{n_i \times n_j}$ for $(i, j) = (1,2)\ or\ (2,1)$ with edge weights equal to the cosine similarity between the PCs.

2.4. Apply the graph-coarsening mapping kernel to identify cross-species mutual nearest neighborhoods.

To increase the stringency and confidence of mapping, we only rely on cells that are *mutual* nearest cross-species neighbors, which are typically defined as two cells reciprocally connected to one another (Haghverdi et al., 2018). However, due to the noise in cell-cell correlations and stochasticity in the kNN algorithms, cross-species neighbors are often randomly assigned from a pool of cells that appear equally similar, decreasing the likelihood of mutual connectivity between individual cells even if they have similar expression profiles. To overcome this limitation, we integrate information from each cell's local neighborhood to establish more robust mutual connectivity between cells across species. Two cells are thus defined as mutual nearest cross-species neighbors when their respective neighborhoods have mutual connectivity.

Specifically, the nearest neighbor graphs N_i calculated in step 1.2 are used to calculate the neighbors of cells t_i hops away along outgoing edges: $\bar{N}_i = {N_i}^{t_i}$, where $\bar{N}_i$ are adjacency matrices that contain the number of paths connecting two cells t_i hops away, for $i = 1\ or\ 2$. t_i determines the length-scale over which we integrate incoming edges for species i. Its default value is 2 if the dataset size is less than 20,000 cells and 3 otherwise. However, cells within tight clusters may have spurious edges connecting to other parts of the manifold only a few hops away. To avoid integrating neighborhood information outside this local structure, we use the Leiden algorithm (Traag et al., 2019) to cluster the graph and identify a local neighborhood size for each cell (the resolution parameter is set to 3 by default). If cell a belongs to cluster c_a, then its neighborhood size is $l_a = |c_a|$. For each row a in $\bar{N}_i$ we only keep the l_a geodesically closest cells, letting the pruned graph be denoted as $\widehat{N}_i$.

Edges outgoing from cell a_i in species i are encoded in the corresponding row in the adjacency matrix: C_{i,a_i}. We compute the fraction of the outgoing edges from each cell that target the local neighborhood of a cell in the other species: $\tilde{C}_{i,a_ib_j} = \sum_{c\in X_{j,b_j}} C_{i,a_ic}$, where X_{j,b_j} is the set of cells in the neighborhood of cell b_j in species j and $\tilde{C}_{i,a_ib_j}$ is the fraction of outgoing edges from cell a_i in species i targeting the neighborhood of cell b_j in species j.

To reduce the density of $\tilde{C}_i$ so as to satisfy computational memory constraints, we remove edges with weight less than 0.1. Finally, we apply the mutual nearest neighborhood criterion by taking the element-wise, geometric mean of the two directed bipartite graphs: $\tilde{C} = \sqrt{\tilde{C}_1 \circ \tilde{C}_2}$. This operation ensures that only bidirectional edges are preserved, as small edge weights in either direction results in small geometric means.

2.5. Assign the k-nearest cross-species neighborhoods for each cell and update edge weights in the gene homology graph.

Given the mutual nearest neighborhoods $\tilde{C} \in \Re^{n_1 \times n_2}$, we select the k nearest neighborhoods for each cell in both directions to update the directed biadjacency matrices C_1 and C_2: $C_1 = KNN(\tilde{C}, k)$ and $C_2 = KNN(\tilde{C}^T \quad, k)$, with $k = 20$ by default.

2.6. Stitch the manifolds.

We use C_1 and C_2 to combine the manifolds N_1 and N_2 into a unified graph. We first weight the edges in N_1 and N_2 to account for the number of shared cross-species neighbors by computing the one-mode projections of C_1 and C_2. In addition, for cells with strong cross-species alignment, we attenuate the weight of their within-species edges. For cells with little to no cross-species alignment, their within-species are kept the same to ensure that the local topological information around cells with no alignment is preserved.

Specifically, we use N_1 and N_2 to mask the edges in the one-mode projections, $\widetilde{N}_1 = U(N_1) \circ (Norm(C_1)Norm(C_2))$ and $\widetilde{N}_2 = U(N_2) \circ (Norm(C_2)Norm(C_1))$, where $U(E)$ sets all edge weights in graph E to 1 and $Norm$ normalizes the outgoing edges from each cell to sum to 1. The minimum edge weight is set to be 0.3 to ensure that neighbors in the original manifolds with no shared cross-species neighbors still retain connectivity: $\widetilde{N}_{1,ij} = min(0.3, \widetilde{N}_{1,ij})$ and $\widetilde{N}_{2,ij} = min(0.3, \widetilde{N}_{2,ij})$ for all edges (i,j). We then scale the within-species edges from cell i by the total weight of its cross-species edges: $\widetilde{N}_{1,i} = (1 - \frac{1}{k}\sum_{j=1}^{n_2} C_{1,ij})\widetilde{N}_{1,i}$ and $\widetilde{N}_{2,i} = (1 - \frac{1}{k}\sum_{j=1}^{n_1} C_{2,ij})\widetilde{N}_{2,i}$. Finally, the within- and cross-species graphs are stitched together to form the combined nearest neighbor graph N: $N = [\widetilde{N}_1 \oplus C_1] \oplus [C_2 \oplus \widetilde{N}_2]$. The overall alignment score between species 1 and 2 is defined as $S = \frac{1}{n_1+n_2}(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} C_{1,ij} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_1} C_{2,ij})$.

3. Gene-gene correlation initialization.

3.1. Update edge weights in the gene-gene bipartite graph with expression correlations.

To compute correlations between gene pairs, we first transfer expressions from one species to the other: $\bar{Z}_{i,n_i m_j} = C_{i,n_i} Z_{j,m_j}$, where $\bar{Z}_{i,n_i m_j}$ is the imputed expressions of gene m_j from species j for cell n_i in species i, and C_{i,n_i} is row n_i of the biadjacency matrix encoding the cross-species neighbors of cell n_i in species i, all for $(i,j) = (1,2)$ and $(2,1)$. We similarly use the manifolds constructed by SAM to smooth the within-species gene expressions using kNN averaging: $\bar{Z}_{j,m_j} = N_{j,m_j} Z_{j,m_j}$, where N_j is the nearest-neighbor graph for species j. We then concatenate the within- and cross-species gene expressions such that the expression of gene m_j from species j in both species is $\bar{Z}_{m_j} = \bar{Z}_{i,m_j} \oplus \bar{Z}_{j,m_j}$.

For all gene pairs in the unpruned homology graph generated in step 1.1., $\hat{G}$, we compute their correlations, $\hat{G}_{ab} := \theta(0)Corr(\bar{Z}_a, \bar{Z}_b)$, where $\theta(0)$ is a Heaviside step function centered at 0 to set negative correlations to zero. We then use the expression correlations to update the corresponding edge weights in $\hat{G}$, which are again normalized through $\hat{G}_{ab} = 0.5 + 0.5\, tanh\,(10\hat{G}_{ab}/\max_b(\hat{G}_{ab}) - 5)$.

Annotation of cell atlases

To annotate the primary zebrafish and *Xenopus* cell types, the cell subtype annotations provided by the original publications (Briggs et al., 2018; Wagner et al., 2018) are coarsened using a combination of the manual matching and developmental hierarchies. For example, as “heart - mature”, “heart - hoxd9a”, “heart”, and “heart field” in zebrafish are all manually matched to “cardiac mesoderm” in *Xenopus*, we label these cells as “heart”. In cases where the matching is insufficient to coarsen the annotations, we use the provided developmental trees to name a group of terminal cell subtypes by their common ontogenic ancestor. The annotations provided by their respective studies were used to label the cells in the *Spongilla*, *Hydra*, and mouse atlases. To annotate the schistosome cells, we used known marker genes to annotate the main schistosome tissue types (Li et al., 2020). Annotations for all single cells in all datasets are provided in **Supplementary Table 1**.

Visualization

The combined manifold N is embedded into 2D projections using UMAP implemented in the scanpy package (Wolf et al., 2018) by *scanpy.tl.umap* with the parameter *min_dist* = 0.1. The sankeyD3 package (https://rdrr.io/github/fbreitwieser/sankeyD3/man/sankeyD3-package.html) in R is used to generate the sankey plots. Edge thickness corresponds to the alignment score between mapped cell types. The alignment score between cell types a and b is defined as $s_{ab} = \frac{1}{|c_a|+|c_b|}\left(\sum_{i\in c_a}\sum_{j\in c_b} C_{1,ij} + \sum_{i\in c_b}\sum_{j\in c_a} C_{2,ij}\right)$, where c_a and c_b are the set of cells in cell types a and b, respectively. Cell type pairs with alignment score less than z are filtered out. By default, z is set to be 0.1. Cell types that did not cluster properly in their respective manifolds were omitted from the sankey plot. In the zebrafish-*Xenopus* comparison, we excluded heart, germline, and olfactory placode cells from both species because they did not cluster in the *Xenopus* atlas. Similarly, the iridoblast, epiphysis, *nanog*+, apoptotic-like, and forerunner cells were excluded because they did not cluster in the zebrafish atlas.

The network graphs in **Figure 4D** are generated using the *networkx* package (https://networkx.github.io) in python. To focus on densely connected cell type groups, we filter out cell type pairs with alignment score less than 0.05.

Identification of gene pairs that drive cell type mappings

We define g_1 and g_2 to contain SAMap-linked genes from species 1 and 2, respectively. Note that a gene may appear multiple times as SAMap allows for one-to-many homology. Let $X_{a_1b_2}$ denote the set of all cells with cross species edges between cell types a_1 and b_2. We calculate the average standardized expression of all cells from species i that are in $X_{a_1b_2}$: $Y_{i,g_i} = \frac{1}{|\{x, x\in X_{a_1b_2}\}|}\sum_{x\in X_{a_1b_2}} \tilde{Z}_{i,x,g_i}$, where $\tilde{Z}_{i,x,g_i} \in \mathfrak{R}^{|g_i|}$ is the standardized expression of genes g_i in cell x. The correlation between Y_{1,g_1} and Y_{2,g_2} can be written as $Corr(Y_{1,g_1}, Y_{2,g_2}) = \sum_{j=1}^{|g_1|} S(Y_{1,g_1})_j \circ S(Y_{2,g_2})_j$, where $S(Z)$ standardizes vector Z to have zero mean and unit variance. We use the summand to identify gene pairs that contribute most positively to the correlation. We assign each gene pair a score: $h_g = T(S(Y_{1,g_1})) \circ T(S(Y_{2,g_2}))$, where $T(Z)$ sets negative values in vector Z to zero in order to ignore lowly-expressed genes. To be inclusive, we begin with the top 1,000 gene pairs according to h_g and filter out gene pairs in which one or both of the genes are not differentially expressed in their respective cell types (p-value > 10^{-2}), have less than 0.2 SAM weight, or are expressed in fewer than 5% of the cells in the cluster. The differential expression of each gene in each cell type is calculated using the Wilcoxon rank-sum test implemented in the *scanpy* function *scanpy.tl.rank_genes_groups*.

Orthology group assignment

We used the Eggnog mapper (v5.0) (Huerta-Cepas et al., 2019) to assign each gene to an orthology group with default parameters. For the zebrafish-to-*Xenopus* mapping, genes are considered paralogs if they map to the same eukaryotic orthology group and orthologs if they map to the same vertebrate orthology group. For the pan-species analysis, we group genes from all species with overlapping orthology assignments. In **Figure 5B**, each column corresponds to one of these groups. As each group may contain multiple genes from each species, we present the expression of the gene with the highest enrichment score per species. All gene names and corresponding orthology groups are reported in **Supplementary Table 4**.

Phylogenetic reconstruction of gene trees

We generate gene trees to validate the identity of genes involved in putative examples of paralog substitution and of *Fox* and *Csrp* transcriptional regulators that are identified as enriched in contractile cells. For this, we first gather protein sequences from potential homologs using the eggnog version 5.0 orthology database (Huerta-Cepas et al., 2019). For the *Fox* and *Csrp* phylogenies, we include all Fox clade I (Larroux et al., 2008) and Csrp/Crip homologs, respectively, from the seven species included in our study.

Alignment of protein sequences is performed with Clustal Omega version 1.2.4 using default settings as implemented on the EMBL EBI web services platform (Madeira et al., 2019). Maximum likelihood tree reconstruction is performed using IQ-TREE version 1.6.12 (Nguyen et al., 2015) with the ModelFinder Plus option (Kalyaanamoorthy et al., 2017). For the *Csrp* tree, we perform 1,000 nonparametric bootstrap replicates to assess node support. For *Fox*, we utilize the ultrafast bootstrap support option with 1,000 replicates. For each gene tree we choose the model that minimizes the Bayesian Information Criterion (BIC) score in ModelFinder. This results in selection of the following models: DCMut+R4 (*Csrp*) and VT+F+R5 (*Fox*). The final consensus trees are visualized and rendered using the ete3 v3.1.1 python toolkit (Huerta-Cepas et al., 2016) and the Interactive Tree of Life v4 (Letunic & Bork, 2019).

KOG functional annotation and enrichment analysis

Using the eggnog mapper, KOG functional annotations are transferred to individual transcripts from their assigned orthology group. For enrichment analysis, all genes enriched in the set of cell type pairs of interest are lumped to form the target set for each species. For example, the target set for *Spongilla* archaeocytes used in **Figure 5C** is composed of all genes enriched between *Spongilla* archaeocytes and other invertebrate stem cells. Note that this set includes genes from other species that are linked by SAMap to the *Spongilla* archeocyte genes. We include genes from other species in the target set to account for differences in KOG functional annotation coverage between species. As such, the annotated transcripts from all 7 species are combined to form the background set. We used a hypergeometric statistical test (Eden et al., 2009) to measure the enrichment of the KOG terms in the target genes compared to the background genes.

Mapping zebrafish and xenopus atlases using existing methods

For benchmarking, we used vertebrate orthologs as determined by Eggnog as input to Harmony (Korsunsky et al., 2019), Liger (Welch et al., 2019), Seurat (Stuart et al., 2019), Scanorama (Hie et al., 2019), BBKNN (Polański et al., 2019), which are all run with default parameters. One-to-one orthologs were selected from one-to-many and many-to-many orthologs by using the bipartite maximum weight matching algorithm implemented in *networkx*. When using the one-to-one orthologs as input for SAMap, we ran for only one iteration. The resulting integrated lower-dimensional coordinates (PCs for Seurat, Harmony, and Scanorama and non-negative matrix factorization coordinates for Liger) and stitched graph (BBKNN and SAMap) were all projected into 2D with UMAP (**Figure 2 – figure supplement 2A**). The integrated coordinates are used to generate a nearest neighbor graph using the correlation distance metric, which is then used to compute the alignment scores in **Figure 2 – figure supplement 2B**. The alignment scores for SAMap and BBKNN are directly computed from their combined graphs.

In-situ hybridization in schistosomes

S. mansoni (strain: NMRI) juveniles are retrieved from infected female Swiss Webster mice (NR-21963) at ~3 weeks post-infection by hepatic portal vein perfusion using 37°C DMEM supplemented with 5 % heat inactivated FBS. The infected mice are provided by the NIAID Schistosomiasis Resource Center for distribution through BEI Resources, NIH-NIAID Contract HHSN272201000005I. In adherence to the Animal Welfare Act and the Public Health Service Policy on Humane Care and Use of Laboratory Animals, all experiments with and care of mice are performed in accordance with protocols approved by the Institutional Animal Care and Use Committees (IACUC) of Stanford University (protocol approval number 30366). *In situ* hybridization experiments are performed as described previously (Tarashansky et al., 2019), using riboprobes synthesized from gene fragments cloned with the listed primers: collagen (Smp_170340): GGTGAAGAAGGCTGTTGTGG, ACGATCCCCTTTCACTCCTG; tropomyosin (Smp_031770): AAGCTGAAGTCGCCTCACTA, CATATGCCTCTTCACGCTGG; troponin (Smp_018250): CGTAAACCTGGTCAGAAGCG, ATCCTTTTCCTCCAGAGCGT; myosin regulatory light chain (Smp_132670): GAGACAGCGAGTAGTGGAGG, TGCCTTCTTTGATTGGAGCT; wnt (Smp_156540): TGTGGTGATGAAGATGGCAG, CCACGGCCACAACACATATT; frizzled (Smp_174350): CGAACAGGCGCATGACAATA, TGCTAGTCCTGTTGTCGTGT.

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 and an NIH grant (1R35GM138061) to BW.

References

Alié, A., Hayashi, T., Sugimura, I., Manuel, M., Sugano, W., Mano, A., Satoh, N., Agata, K., & Funayama, N. (2015). The ancestral gene repertoire of animal stem cells. *Proceedings of the National Academy of Sciences*, *112*(51), E7093–E7100. https://doi.org/10.1073/pnas.1514789112

Arendt, D., Bertucci, P. Y., Achim, K., & Musser, J. M. (2019). Evolution of neuronal types and families. *Current Opinion in Neurobiology*, *56*, 144–152. https://doi.org/10.1016/j.conb.2019.01.022

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., & Wagner, G. P. (2016). The origin and evolution of cell types. *Nature Reviews Genetics*, *17*(12), 744–757. https://doi.org/10.1038/nrg.2016.127

Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., & Kharchenko, P. V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nature Methods*, *16*(8), 695–698. https://doi.org/10.1038/s41592-019-0466-z

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, *3*(4), 346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W., & Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, *360*(6392), eaar5780. https://doi.org/10.1126/science.aar5780

Brunet, T., Fischer, A. H., Steinmetz, P. R., Lauri, A., Bertucci, P., & Arendt, D. (2016). The evolutionary origin of bilaterian smooth and striated myocytes. *eLife*, *5*, e19607. https://doi.org/10.7554/eLife.19607

Buzgariu, W., Al Haddad, S., Tomczyk, S., Wenger, Y., & Galliot, B. (2015). Multi-functionality and plasticity characterize epithelial cells in *Hydra*. *Tissue Barriers*, *3*(4), e1068908. https://doi.org/10.1080/21688370.2015.1068908

Cao, C., Lemaire, L. A., Wang, W., Yoon, P. H., Choi, Y. A., Parsons, L. R., Matese, J. C., Wang, W., Levine, M., & Chen, K. (2019). Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature*, *571*(7765), 349–354. https://doi.org/10.1038/s41586-019-1385-y

Dubaissi, E., Rousseau, K., Lea, R., Soto, X., Nardeosingh, S., Schweickert, A., Amaya, E., Thornton, D. J., & Papalopulu, N. (2014). 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*(7), 1514–1525. https://doi.org/10.1242/dev.102426

Eddy, S. R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Computational Biology*, *4*(5), e1000069. https://doi.org/10.1371/journal.pcbi.1000069

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. https://doi.org/10.1186/1471-2105-10-48

El-Brolosy, M. A., Kontarakis, Z., Rossi, A., Kuenne, C., Günther, S., Fukuda, N., Kikhi, K., Boezio, G. L. M., Takacs, C. M., Lai, S.-L., Fukuda, R., Gerri, C., Giraldez, A. J., & Stainier, D. Y. R. (2019). Genetic compensation triggered by mutant mRNA degradation. *Nature*, *568*(7751), 193–197. https://doi.org/10.1038/s41586-019-1064-z

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, *360*(6391), eaaq1736. https://doi.org/10.1126/science.aaq1736

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, *14*(5), 360–366. https://doi.org/10.1038/nrg3456

Geirsdottir, L., David, E., Keren-Shaul, H., Weiner, A., Bohlen, S. C., Neuber, J., Balic, A., Giladi, A., Sheban, F., Dutertre, C.-A., Pfeifle, C., Peri, F., Raffo-Romero, A., Vizioli, J., Matiasek, K., Scheiwe, C., Meckel, S., Mätz-Rensing, K., van der Meer, F., … Prinz, M. (2019). Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program. *Cell*, *179*(7), 1609-1622.e16. https://doi.org/10.1016/j.cell.2019.11.010

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427. https://doi.org/10.1038/nbt.4091

Hie, B., Bryson, B., & Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, *37*(6), 685–691. https://doi.org/10.1038/s41587-019-0113-3

Hu, M., Zheng, X., Fan, C.-M., & Zheng, Y. (2020). Lineage dynamics of the endosymbiotic cell type in the soft coral Xenia. *Nature*, *582*(7813), 534–538. https://doi.org/10.1038/s41586-020-2385-7

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Larroux, C., Luke, G. N., Koopman, P., Rokhsar, D. S., Shimeld, S. M., & Degnan, B. M. (2008). Genesis and Expansion of Metazoan Transcription Factor Gene Classes. *Molecular Biology and Evolution*, *25*(5), 980–996. https://doi.org/10.1093/molbev/msn047

Laumer, C. E., Hejnol, A., & Giribet, G. (2015). Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *eLife*, *4*, e05503. https://doi.org/10.7554/eLife.05503

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239

Li, P., Sarfati, D. N., Xue, Y., Yu, X., Tarashansky, A. J., Quake, S. R., & Wang, B. (2020). *Single-cell analysis of* Schistosoma mansoni *reveals a conserved genetic program controlling germline stem cell fate* [Preprint]. https://doi.org/10.1101/2020.07.06.190033

Littlewood, D. T. J., & Waeschenbach, A. (2015). Evolution: A Turn Up for the Worms. *Current Biology*, *25*(11), R457–R460. https://doi.org/10.1016/j.cub.2015.04.012

Madeira, F., Park, Y. mi, Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Miles, L. B., Darido, C., Kaslin, J., Heath, J. K., Jane, S. M., & Dworkin, S. (2017). Mis-expression of grainyhead-like transcription factors in zebrafish leads to defects in enveloping layer (EVL) integrity, cellular morphogenesis and axial extension. *Scientific Reports*, *7*(1), 17607. https://doi.org/10.1038/s41598-017-17898-7

Musser, J. M., Schippers, K. J., Nickel, M., Mizzon, G., Kohn, A. B., Pape, C., Hammel, J. U., Wolf, F., Liang, C., Hernández-Plaza, A., Achim, K., Schieber, N. L., Francis, W. R., Vargas R., S., Kling, S., Renkert, M., Feuda, R., Gaspar, I., Burkhardt, P., … Arendt, D. (2019). *Profiling cellular diversity in sponges informs animal cell type and nervous system evolution* [Preprint]. https://doi.org/10.1101/758276

Nehrt, N. L., Clark, W. T., Radivojac, P., & Hahn, M. W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, *7*(6), e1002073. https://doi.org/10.1371/journal.pcbi.1002073

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C., & Göttgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, *566*(7745), 490–495. https://doi.org/10.1038/s41586-019-0933-9

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C., & Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, *360*(6391), eaaq1723. https://doi.org/10.1126/science.aaq1723

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A., & Park, J.-E. (2019). BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics*, btz625. https://doi.org/10.1093/bioinformatics/btz625

Prince, V. E., & Pickett, F. B. (2002). Splitting pairs: The diverging fates of duplicated genes. *Nature Reviews Genetics*, *3*(11), 827–837. https://doi.org/10.1038/nrg928

Reddien, P. W. (2018). The Cellular and Molecular Basis for Planarian Regeneration. *Cell*, *175*(2), 327–345. https://doi.org/10.1016/j.cell.2018.09.021

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Human Cell Atlas Meeting Participants. (2017). The Human Cell Atlas. *eLife*, *6*, e27041. https://doi.org/10.7554/eLife.27041

Scimone, M. L., Cote, L. E., & Reddien, P. W. (2017). Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature*, *551*(7682), 623–628. https://doi.org/10.1038/nature24660

Sebé-Pedrós, A., Chomsky, E., Pang, K., Lara-Astiaso, D., Gaiti, F., Mukamel, Z., Amit, I., Hejnol, A., Degnan, B. M., & Tanay, A. (2018). Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nature Ecology & Evolution*, *2*(7), 1176–1188. https://doi.org/10.1038/s41559-018-0575-6

Shafer, M. E. R. (2019). Cross-Species Analysis of Single-Cell Transcriptomic Data. *Frontiers in Cell and Developmental Biology*, *7*, 175. https://doi.org/10.3389/fcell.2019.00175

Siebert, S., Farrell, J. A., Cazet, J. F., Abeykoon, Y., Primack, A. S., Schnitzler, C. E., & Juliano, C. E. (2019). Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science*, *365*(6451), eaav9314. https://doi.org/10.1126/science.aav9314

Stamboulian, M., Guerrero, R. F., Hahn, M. W., & Radivojac, P. (2020). The ortholog conjecture revisited: The value of orthologs and paralogs in function prediction. *Bioinformatics*, *36*(Supplement_1), i219–i226. https://doi.org/10.1093/bioinformatics/btaa468

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Studer, R. A., & Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, *25*(5), 210–216. https://doi.org/10.1016/j.tig.2009.03.004

Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R., & Wang, B. (2019). Self-assembling manifolds in single-cell RNA sequencing data. *eLife*, *8*, e48994. https://doi.org/10.7554/eLife.48994

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, 14.

Tosches, M. A., & Arendt, D. (2013). The bilaterian forebrain: An evolutionary chimaera. *Current Opinion in Neurobiology*, *23*(6), 1080–1089. https://doi.org/10.1016/j.conb.2013.09.005

Tosches, M. A., Yamawaki, T. M., Naumann, R. K., Jacobi, A. A., Tushev, G., & Laurent, G. (2018). Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science*, *360*(6391), 881–888. https://doi.org/10.1126/science.aar4237

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S., & Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, *360*(6392), 981–987.

Wang, B., Lee, J., Li, P., Saberi, A., Yang, H., Liu, C., Zhao, M., & Newmark, P. A. (2018). Stem cell heterogeneity drives the parasitic life cycle of Schistosoma mansoni. *eLife*, *7*, e35449. https://doi.org/10.7554/eLife.35449

Weir, K., Dupre, C., van Giesen, L., Lee, A. S.-Y., & Bellono, N. W. (2020). A molecular filter for the cnidarian stinging response. *eLife*, *9*, e57578. https://doi.org/10.7554/eLife.57578

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, *177*(7), 1873-1887.e17. https://doi.org/10.1016/j.cell.2019.05.006

Wendt, G. R., & Collins, J. J. (2016). Schistosomiasis as a disease of stem cells. *Current Opinion in Genetics & Development*, *40*, 95–102. https://doi.org/10.1016/j.gde.2016.06.010

Wendt, G. R., Collins, J. N., Pei, J., Pearson, M. S., Bennett, H. M., Loukas, A., Berriman, M., Grishin, N. V., & Collins, J. J. (2018). Flatworm-specific transcriptional regulators promote the specification of tegumental progenitors in Schistosoma mansoni. *eLife*, *7*, e33221. https://doi.org/10.7554/eLife.33221

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Wong, E., Mölter, J., Anggono, V., Degnan, S. M., & Degnan, B. M. (2019). Co-expression of synaptic genes in the sponge Amphimedon queenslandica uncovers ancient neural submodules. *Scientific Reports*, *9*(1), 15781. https://doi.org/10.1038/s41598-019-51282-x

Yan, K.-K., Wang, D., Rozowsky, J., Zheng, H., Cheng, C., & Gerstein, M. (2014). OrthoClust: An orthology-based network framework for clustering data across multiple species. *Genome Biology*, *15*(8), R100. https://doi.org/10.1186/gb-2014-15-8-r100

Zeng, A., Li, H., Guo, L., Gao, X., McKinney, S., Wang, Y., Yu, Z., Park, J., Semerad, C., Ross, E., Cheng, L.-C., Davies, E., Lei, K., Wang, W., Perera, A., Hall, K., Peak, A., Box, A., & Sánchez Alvarado, A. (2018). Prospectively Isolated Tetraspanin+ Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell*, *173*(7), 1593-1608.e20. https://doi.org/10.1016/j.cell.2018.05.006

Figures

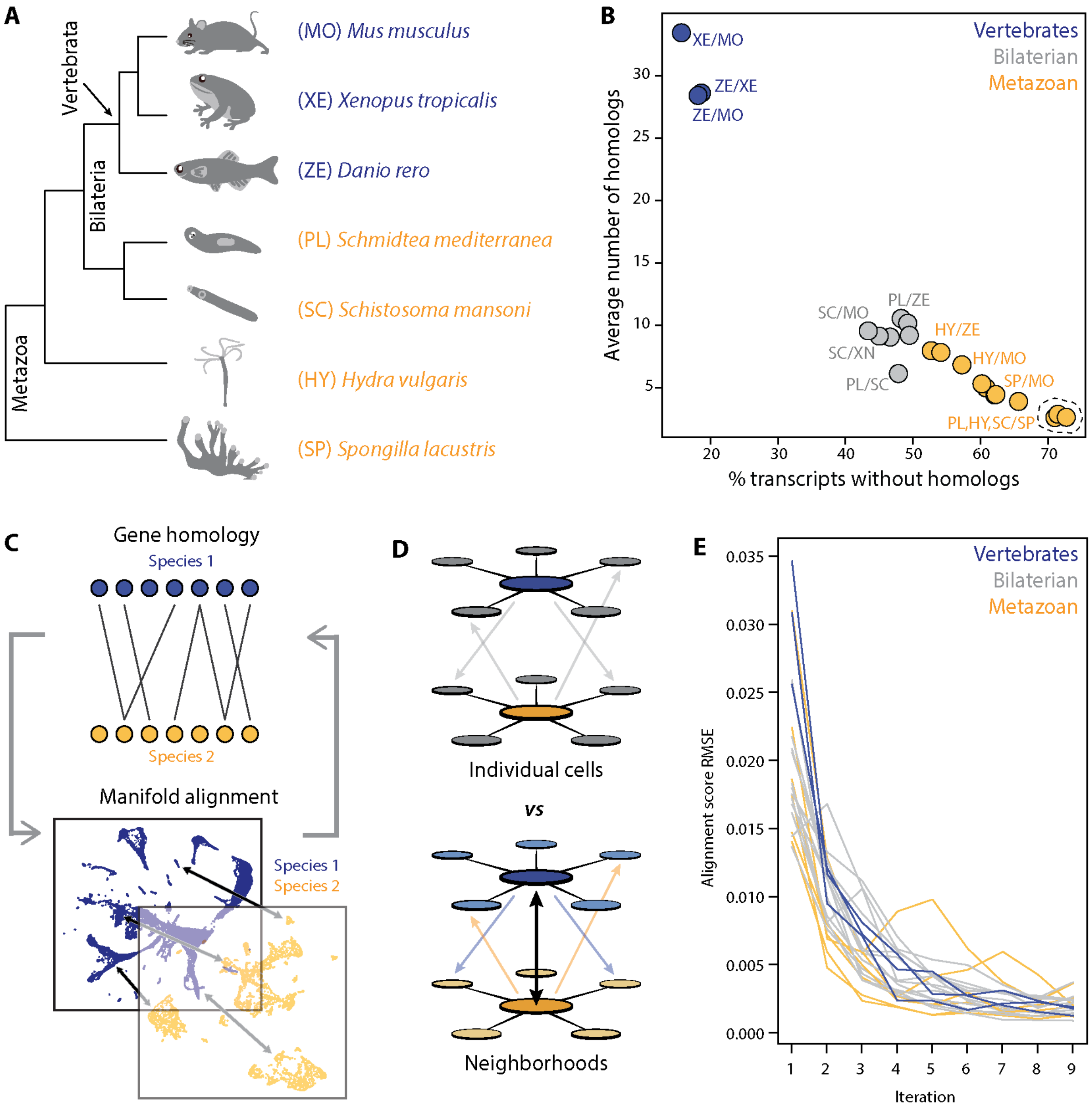

Figure 1: SAMap addresses challenges in mapping cell atlases of distantly related species. (A) Schematic showing the phylogenetic relationships among 7 species analyzed. (B) Challenges in mapping single-cell transcriptomes. Gene duplications cause large numbers of homologs per gene, determined by reciprocal 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. (E) Convergence of SAMap is evaluated by the root mean square error (RMSE) of the alignment scores between mapped clusters in adjacent iterations for all 21 pairwise comparisons of the 7 species.

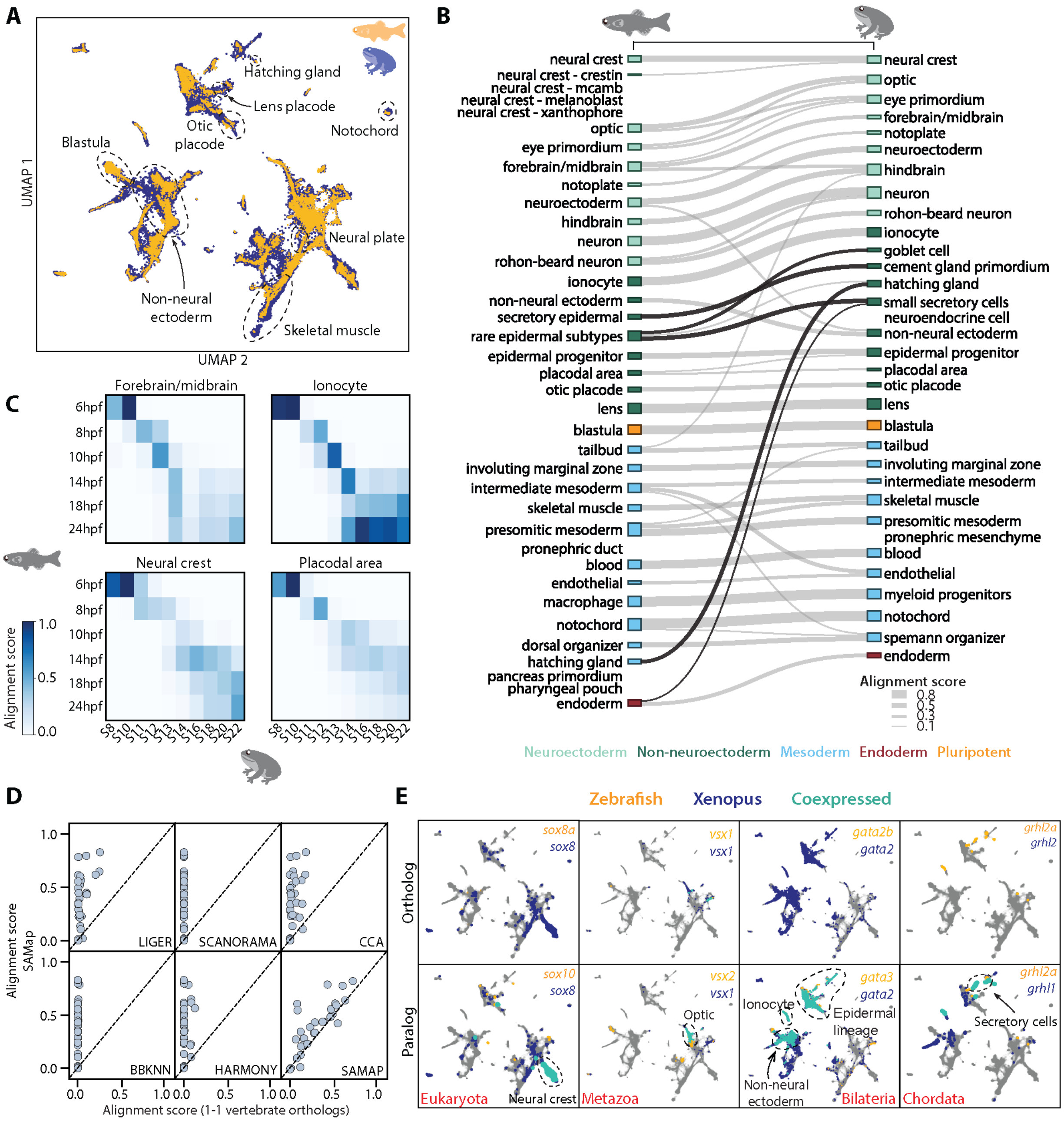

Figure 2: SAMap successfully maps *D. rerio* and *X. tropicalis* atlases and reveals prevalent paralog substitutions. (A) UMAP projection of the combined zebrafish (yellow) and *Xenopus* (blue) manifolds, with example cell types circled. (B) Sankey plot summarizing the cell type mappings. Edges with 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. X-axis: zebrafish. Y-axis: *Xenopus*. (D) SAMap alignment scores compared to those of benchmarking methods using one-to-one vertebrate orthologs as input. Each dot represents a cell type pair supported by ontogeny annotations. (E) 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 **Figure 2 – figure supplement 3**.

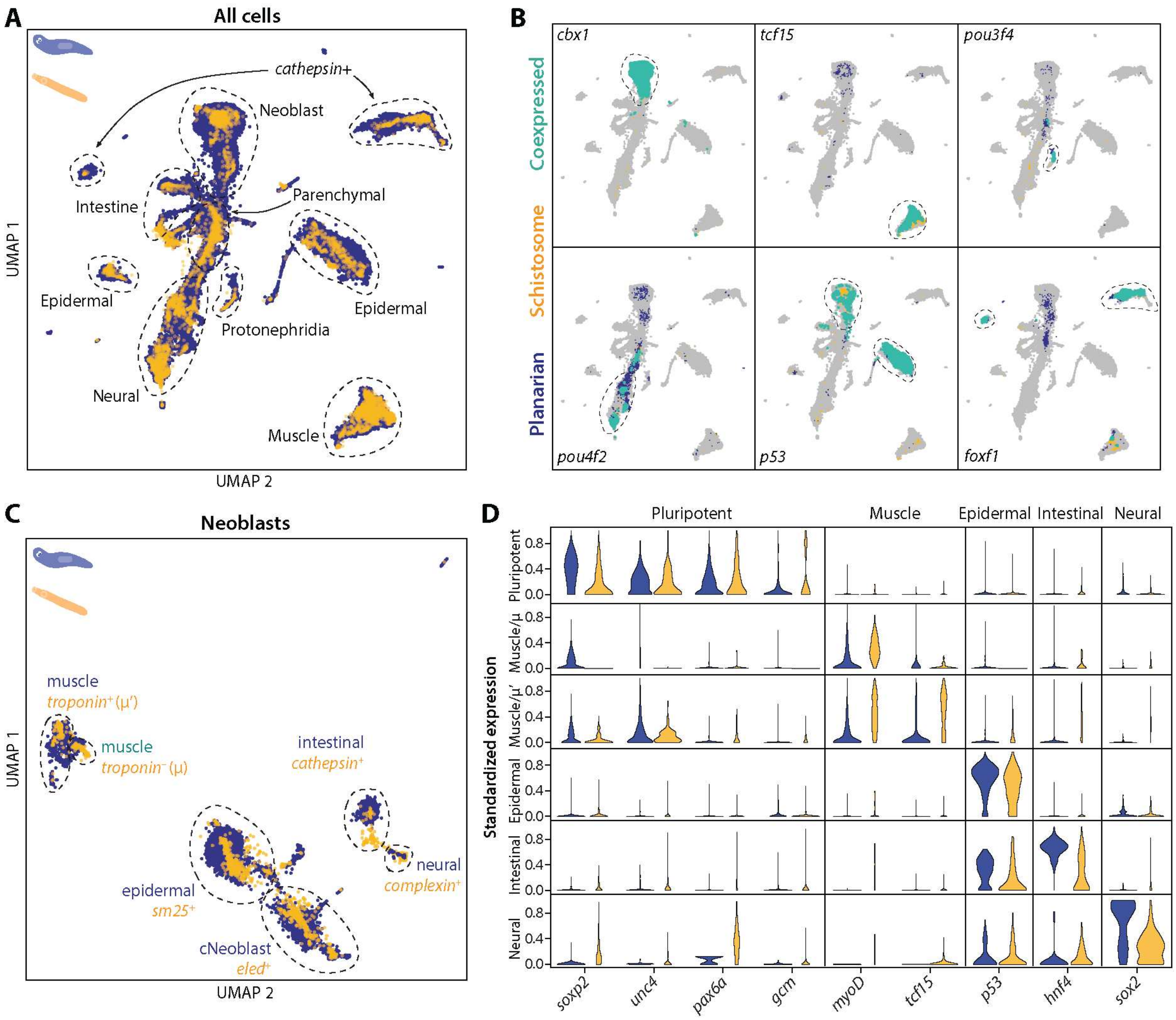

Figure 3: SAMap transfers cell type information from a well-annotated organism (planarian *S. mediterranea*) to its less-studied cousin (schistosome *S. mansoni*) and identifies parallel stem cell compartments. (A) UMAP projection of the combined manifolds. Tissue type annotations are adopted from the *S. mediterranea* atlas (Fincher et al., 2018). The schistosome atlas was collected from juvenile worms, which we found to contain neoblasts with an abundance comparable to that of planarian neoblasts (Li et al., 2020). (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 (Zeng et al., 2018), 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.

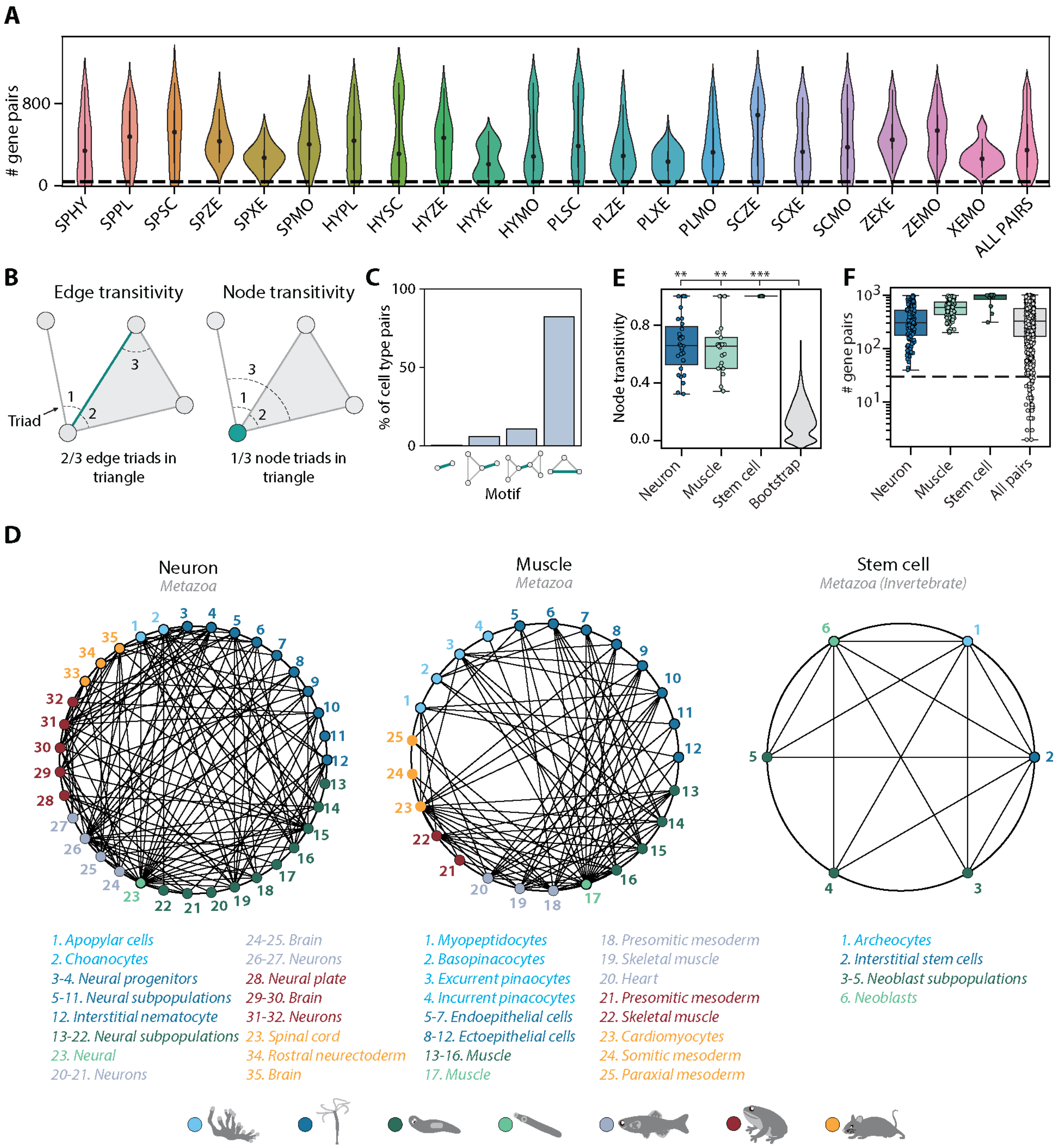

Figure 4: Mapping evolutionarily distant species identifies densely connected cell type groups. (A) Violin plots showing the number of enriched gene pairs in cell type mappings from all 21 pairwise mappings between the 7 species. 87% of cell type mappings have greater than 40 enriched gene pairs (dotted line). 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) The percentage of cell type pairs that are topologically equivalent to the green edge in each illustrated motif. (D) Network graphs showing highly connected cell type families. 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. (E) Boxplot showing the median and interquartile ranges of node transitivities for highly connected cell type groups. For all box plots, the whiskers denote the maximum and minimum observations. 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. $*p < 5x10^{-3}$, $** p < 5x10^{-5}$, $***p < 5x10^{-7}$. (F) Boxplot showing the median and interquartile ranges of the number of enriched gene pairs in highly connected cell type groups. All cell type connections in these groups have at least 40 enriched gene pairs (dotted line).

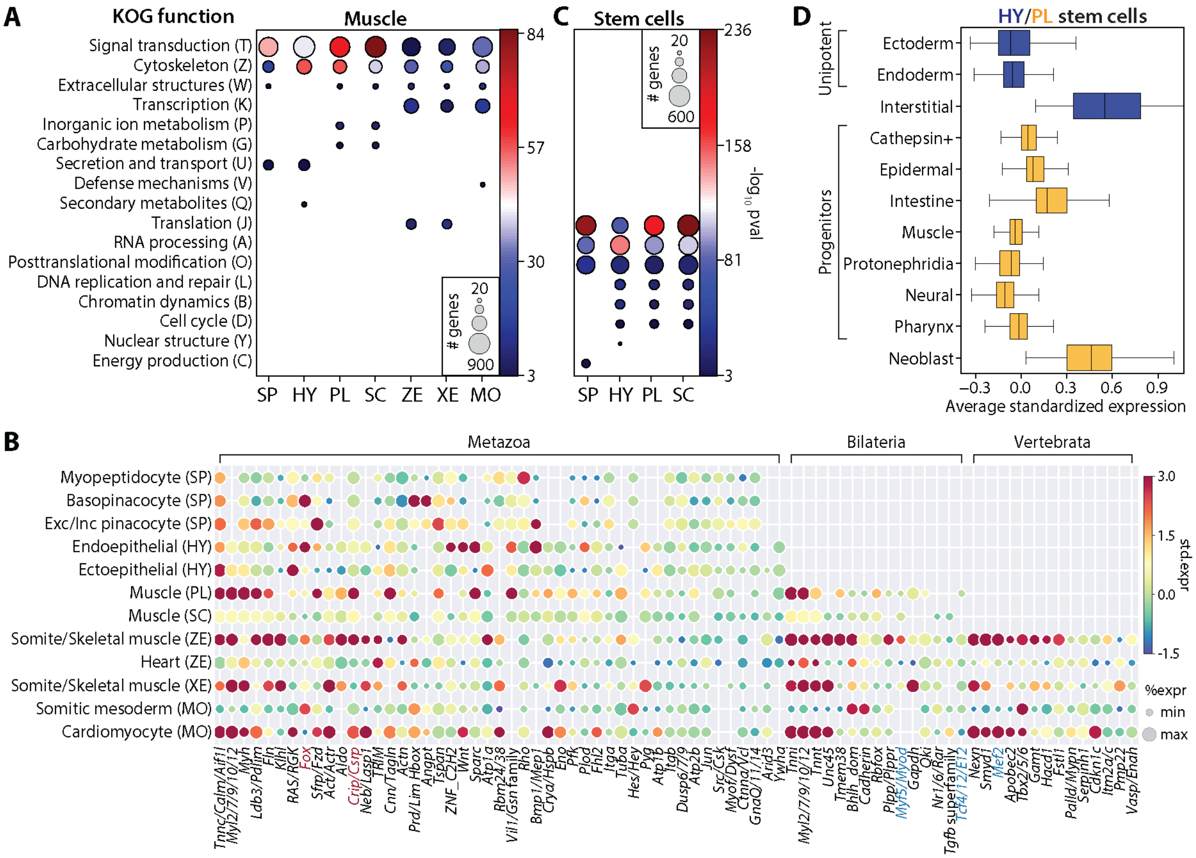

Figure 5: SAMap identifies muscle and stem cell transcriptional signatures conserved across species. (A) Enrichment of KOG functional annotations calculated for genes shared in contractile cell types. For each species, genes enriched in individual contractile cell types are combined. (B) Expression and 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. (Fincher et al., 2018). Neoblasts: cluster 0 in Fincher et al. (Fincher et al., 2018) that does not express any tissue-specific markers.

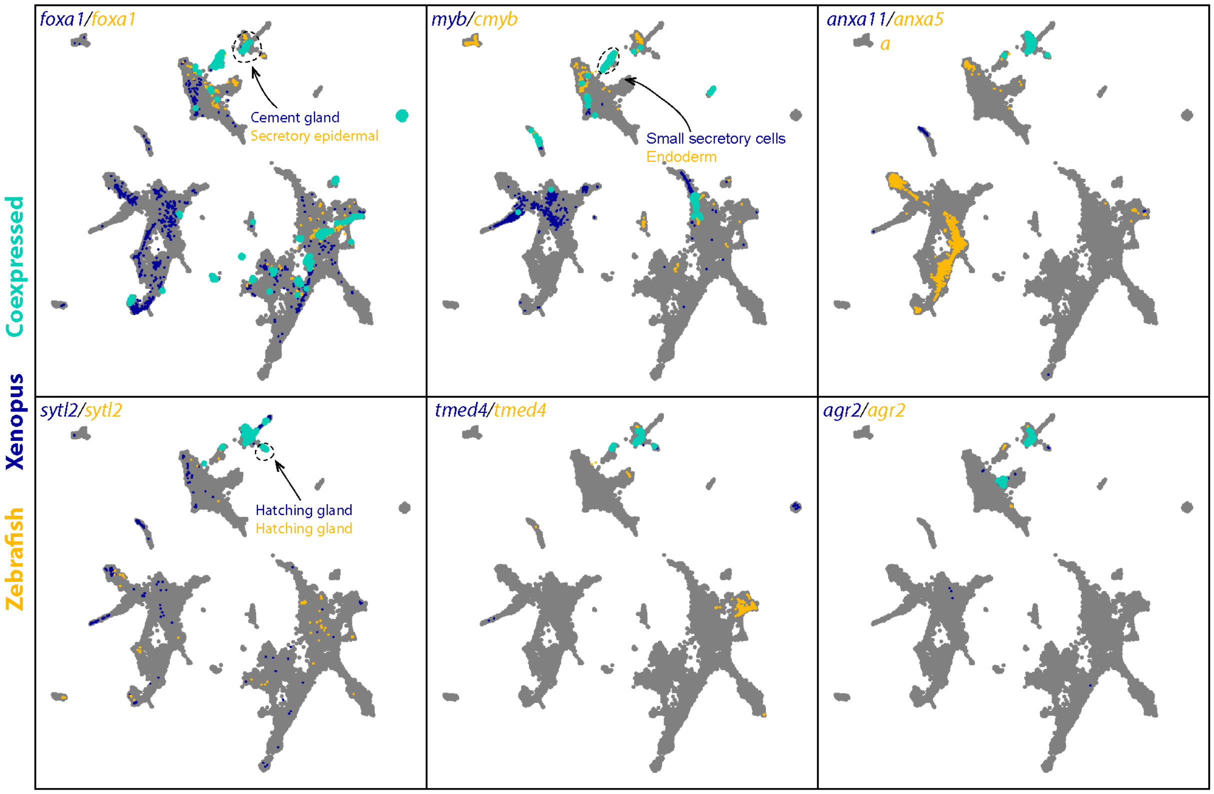

Figure 2 – figure supplement 1: Expression of selected genes enriched in *D. rerio* and *X. tropicalis* secretory cell types. Expressions of orthologous gene pairs linked by SAMap are overlaid on the combined UMAP projection. Expressing cells are color-coded by species, with those connected across species colored cyan. Cells with no expression are shown in gray. The mapped secretory cell types are highlighted with circles.

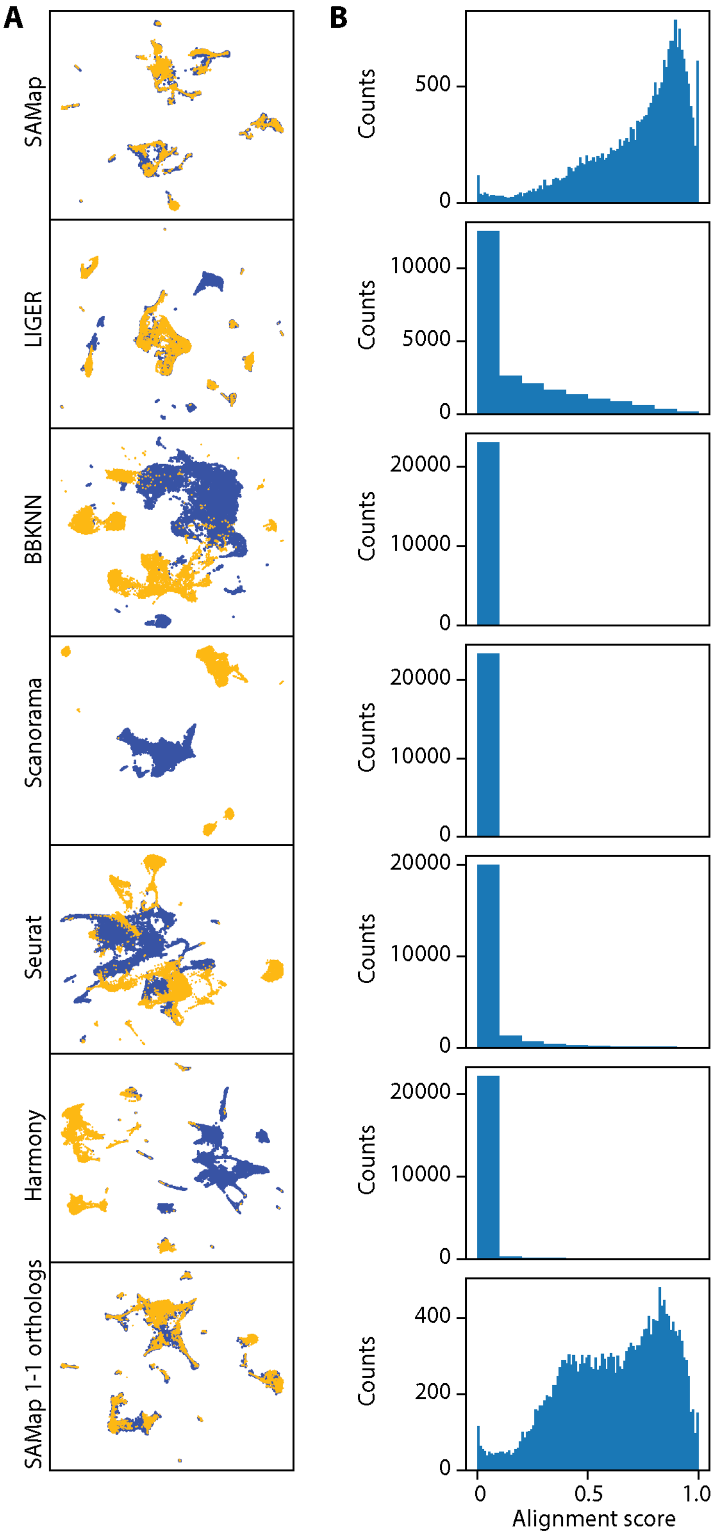

Figure 2 – figure supplement 2: Existing methods failed to map *D. rerio* and *X. tropicalis* atlases. (A) UMAP projections of the integration results from SAMap using the

full homology graph, compared to Liger, BBKNN, Scanorama, Seurat, Harmony, and SAMap using 1-1 orthologs. For fair comparisons, all methods were run on the *D. rerio* and *X. torpicalis* atlases subsampled to approximately 15,000 cells to satisfy computational constraints of Seurat and Liger. (B) Distribution of alignment scores between individual cells.

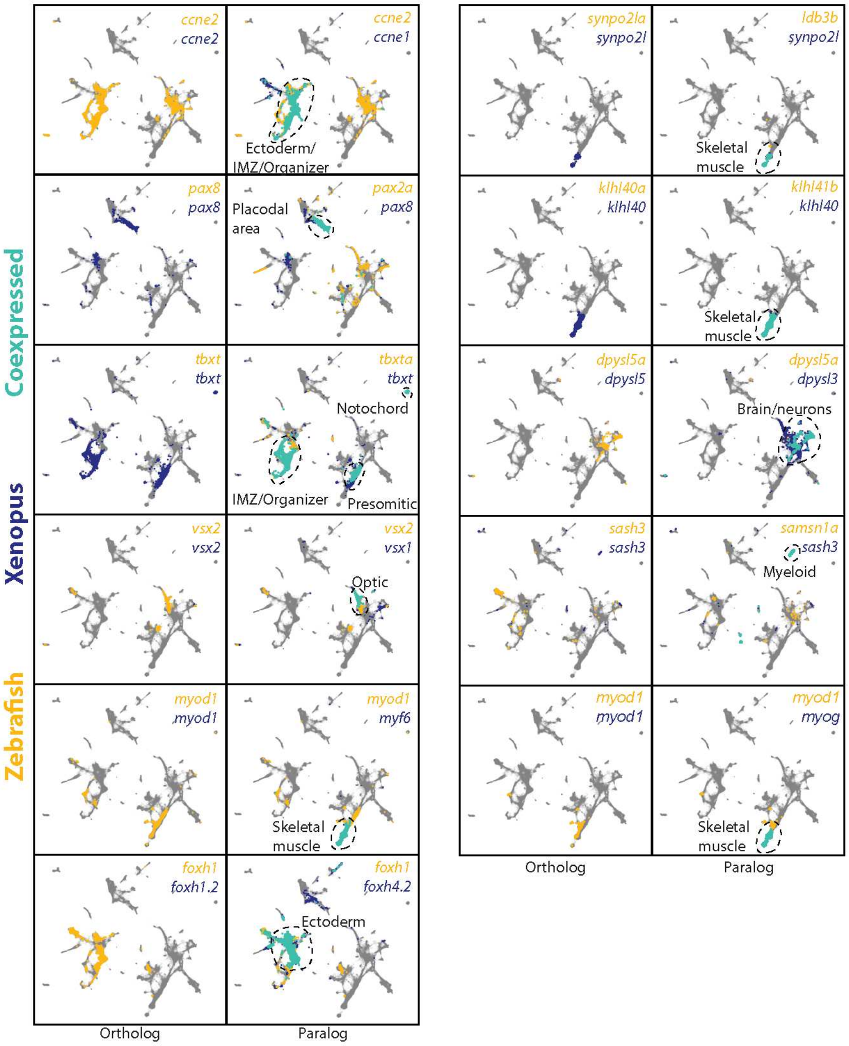

Figure 2 – figure supplement 3: Representative examples of paralog substitution events in *D. rerio* and *X. tropicalis* atlases. Expressions of orthologous and paralogous gene pairs are overlaid on the combined UMAP projection. Expressing cells are color-coded by species, with those that are connected across species colored in cyan. Cells with no expression are shown in gray.

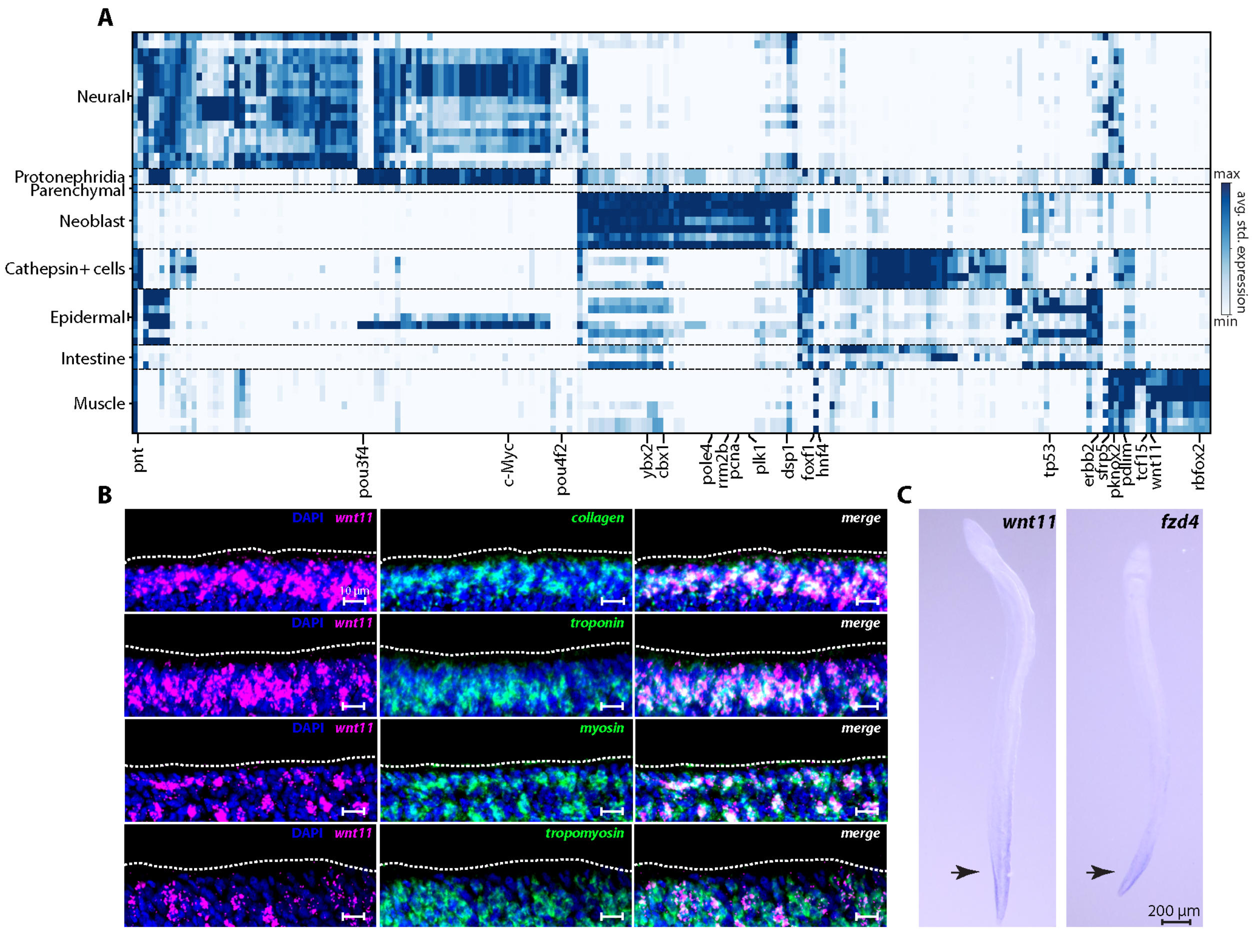

Figure 3 – figure supplement 1: **SAMap-linked gene pairs that are enriched in cell type pairs between *S. mediterranea* and *S. mansoni*.** (A) Rows: linked cell types. Schistosome cell types correspond to leiden clusters. Columns: genes linked by SAMap with overlapping eukaryotic Eggnog orthology groups. 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 (Scimone et al., 2017; Reddien, 2018). The presence of an anterior-posterior expression gradient of *wnt* and *frizzled* in muscles of schistosome juveniles suggests that they may have similar functional roles in patterning during development.

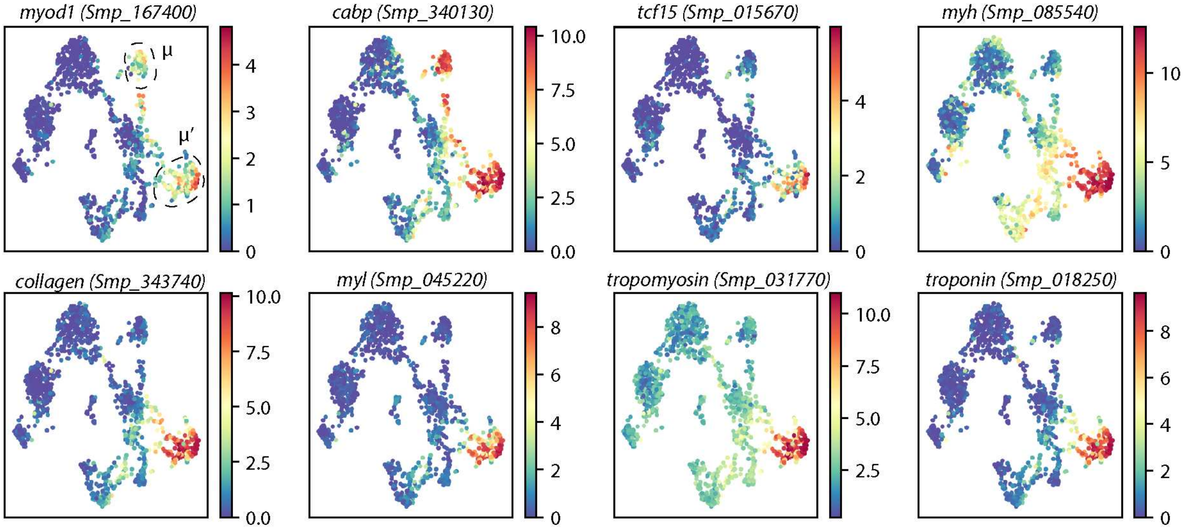

Figure 3 – figure supplement 2: Schistosome neoblasts express canonical muscle markers in muscle progenitors. UMAP projections of schistosome neoblasts with gene expressions overlaid. μ and μ' cells are circled. Colormap: expression in units of $log_2(D + 1)$. For visualization, expression was smoothed 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.

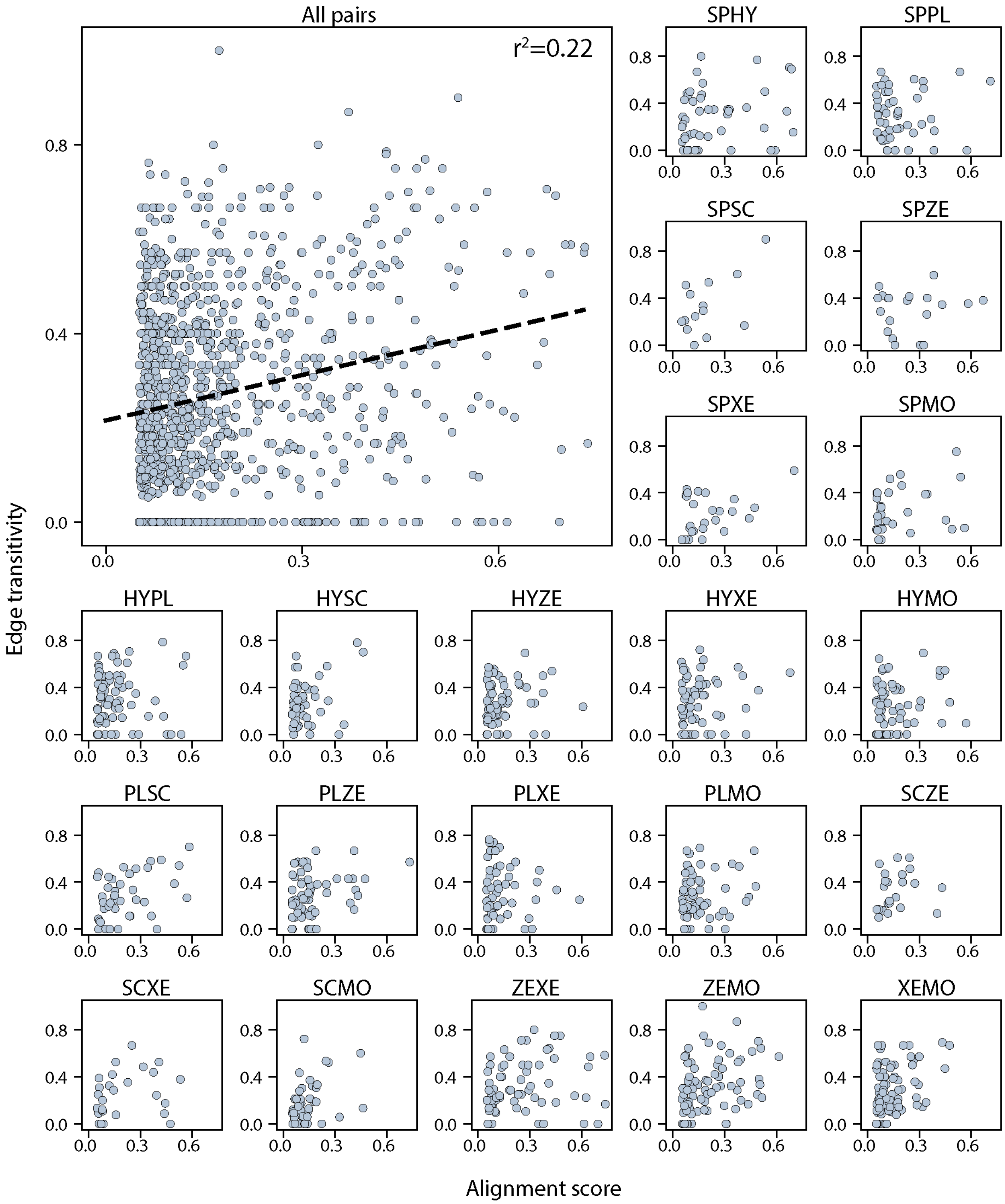

Figure 4 – figure supplement 1: **Alignment scores are mostly independent of edge transitivity.** Top left: alignment scores and edge transitivity for all cell type pairs in the connectivity graph including the 7 species. Dotted line: the linear best fit, with the Pearson

correlation coefficient reported at the top. Alignment scores and edge transitivity for individual species pairs are shown in the remaining subplots.

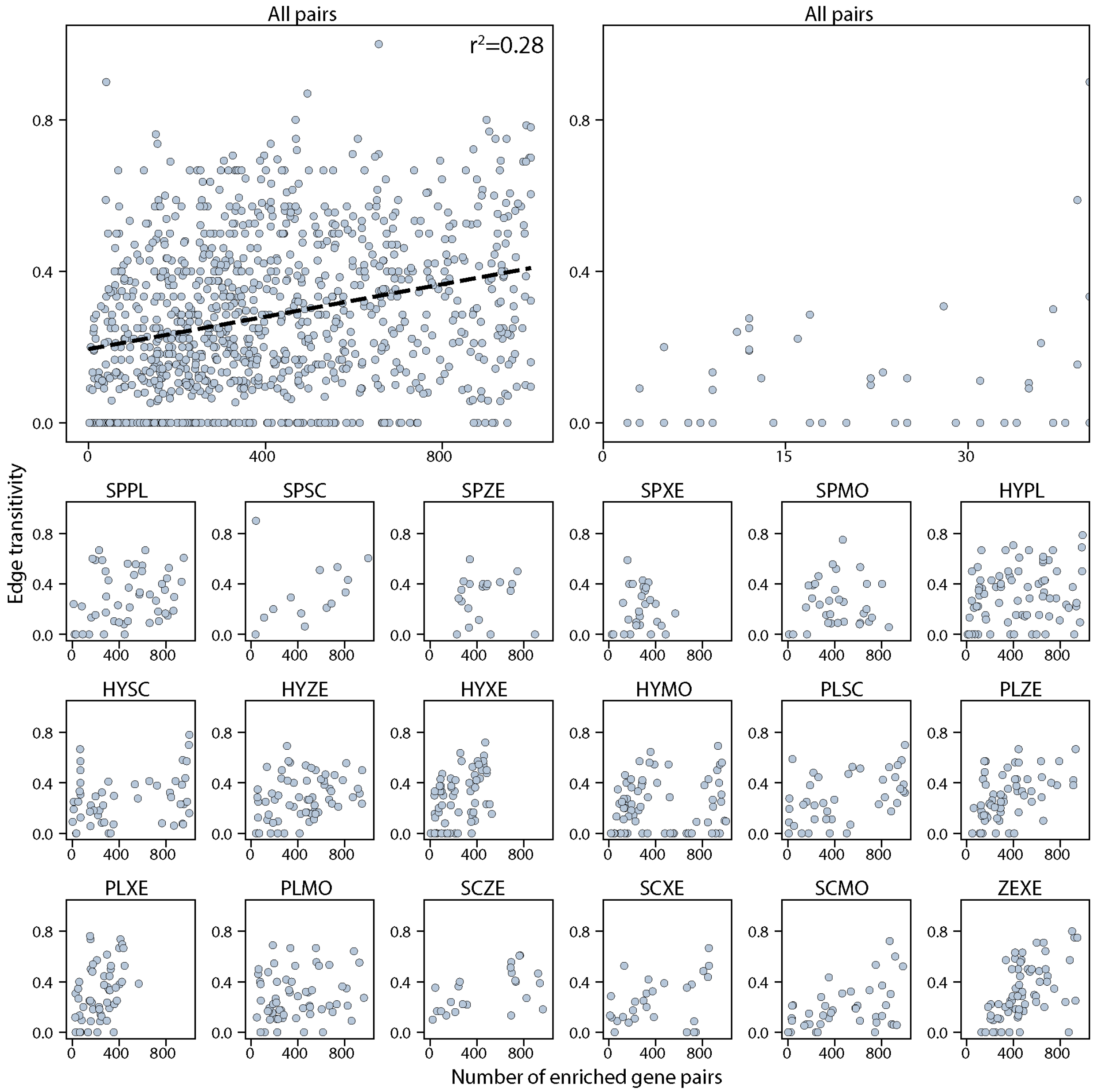

Figure 4 – figure supplement 2: Number of enriched gene pairs are mostly independent of edge transitivity. Top left: The edge transitivity is plotted against the number of enriched gene pairs for all cell type pairs in the connectivity graph. Dotted line: the linear best fit, with the Pearson correlation coefficient reported at the top. Top right: magnified view of the mapped cell type pairs supported by small numbers of gene pairs

(<40) to show those edges have low transitivity scores (<0.4). The sublots below show the number of enriched gene pairs and edge transitivity for individual species pairs.

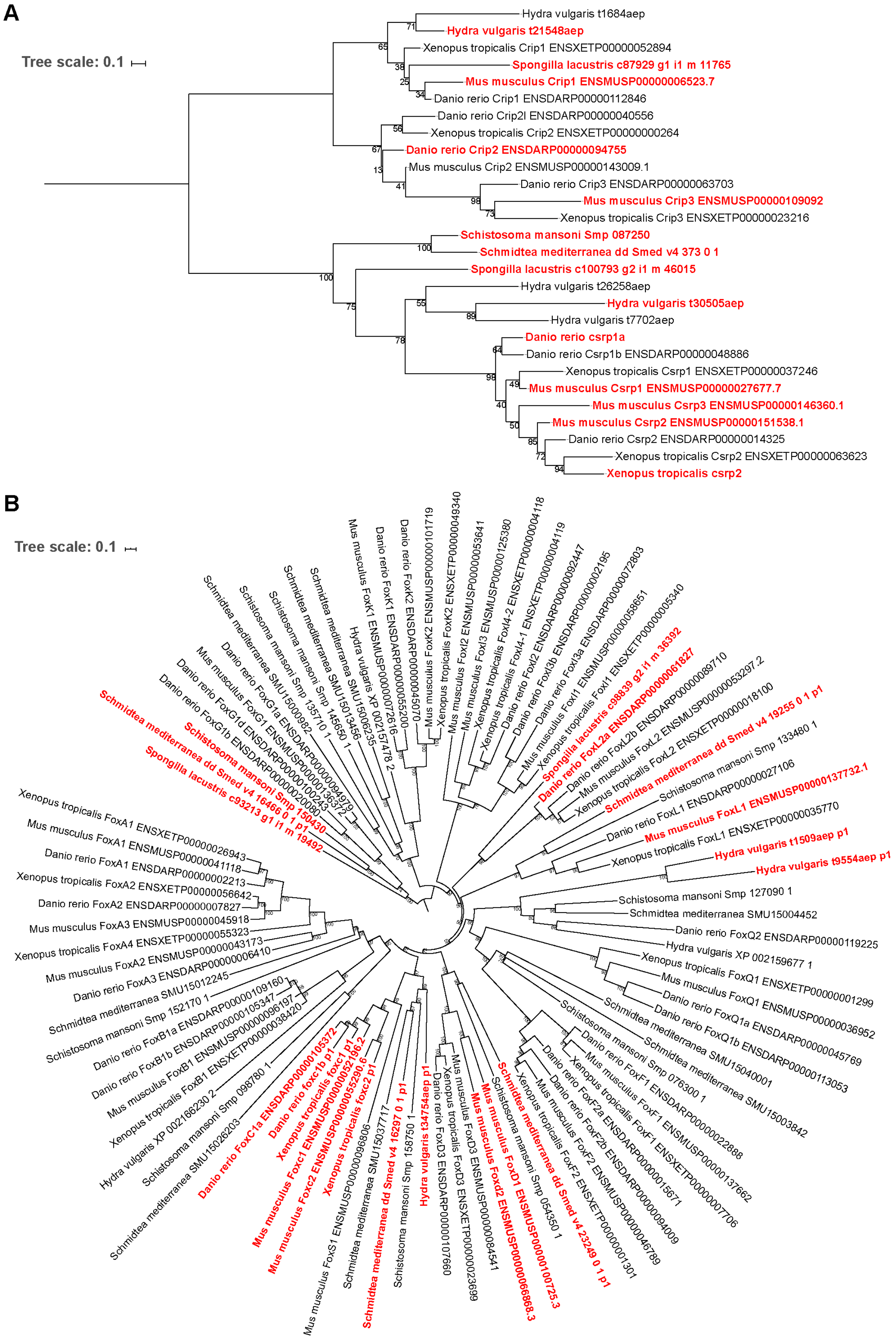
A
Tree scale: 0.1
Hydra vulgaris t1684aep
Hydra vulgaris t21548aep
Xenopus tropicalis Crip1 ENSXETP00000052894
Spongilla lacustris c87929 g1 i1 m 11765
Mus musculus Crip1 ENSMUSP00000006523.7
Danio rerio Crip1 ENSDARP00000112846
Danio rerio Crip2l ENSDARP00000040556
Xenopus tropicalis Crip2 ENSXETP00000000264
Danio rerio Crip2 ENSDARP00000094755
Mus musculus Crip2 ENSMUSP00000143009.1
Danio rerio Crip3 ENSDARP00000063703
Mus musculus Crip3 ENSMUSP00000109092
Xenopus tropicalis Crip3 ENSXETP00000023216
Schistosoma mansoni Smp 087250
Schmidtea mediterranea dd Smed v4 373 0 1
Spongilla lacustris c100793 g2 i1 m 46015
Hydra vulgaris t26258aep
Hydra vulgaris t30505aep
Hydra vulgaris t7702aep
Danio rerio csrp1a
Danio rerio Csrp1b ENSDARP00000048886
Xenopus tropicalis Csrp1 ENSXETP00000037246
Mus musculus Csrp1 ENSMUSP00000027677.7
Mus musculus Csrp3 ENSMUSP00000146360.1
Mus musculus Csrp2 ENSMUSP00000151538.1
Danio rerio Csrp2 ENSDARP00000014325
Xenopus tropicalis Csrp2 ENSXETP00000063623
Xenopus tropicalis csrp2
71
65
38
25
34
58
67
13
41
98
73
100
100
75
55
89
78
64
98
49
40
50
85
72
94
B
Tree scale: 0.1
Mus musculus FoxK1 ENSMUSP00000055200
Danio rerio FoxK1 ENSDARP00000045070
Danio rerio FoxK2 ENSDARP00000101719
Mus musculus FoxK2 ENSMUSP00000049340
Xenopus tropicalis FoxK2 ENSXETP00000053641
Mus musculus FoxI2 ENSMUSP00000125380
Mus musculus FoxI3 ENSMUSP00000004118
Xenopus tropicalis FoxI4-2 ENSXETP00000004119
Xenopus tropicalis FoxI4-1 ENSXETP00000092447
Danio rerio FoxI2 ENSDARP00000002195
Danio rerio FoxI3b ENSDARP00000072803
Danio rerio FoxI3a ENSDARP00000058651
Mus musculus FoxI1 ENSMUSP00000005340
Xenopus tropicalis FoxI1 ENSXETP00000036392
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 ENSMUSP00000062166.2
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
Mus musculus FoxK1 ENSMUSP00000072616

Figure 5 – figure supplement 1: 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: Cell type annotations for the zebrafish-*Xenopus* mapping. Correspondence between the cell type annotations provided in the original study (Briggs et al., 2018; Wagner et al., 2018) and corresponding annotations used in this study is provided for both *D. rerio* and *X. tropicalis* atlases.

Supplementary Table 3: Identified paralogs with greater expression similarity than orthologs 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 whether the paralogs are considered as a paralog substitution (defined as when the substituted ortholog is either absent or lowly-expressed with no cell-type specificity). Highlighted rows are shown in **Figure 2E** and **Figure 2 – figure supplement 3**.

Supplementary Table 4: Genes enriched in contractile cell types and invertebrate stem cells highlighted in Figure 4D. 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 (Musser et al., 2019).

Supplementary Table 5: Cell types in the cell type families shown in Figure 4D. For the schistosome cell types, we annotated two neural clusters, both of which express the neural marker *complexin* (Li et al., 2020). 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 a previous study (Wendt et al., 2018).