Published on Sep 29, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-vw5e-2670

Comparing gene expression across species based on protein structure

We investigated protein structure predictions as an alternative to protein sequence homology for comparing single-cell RNA-seq data across species.

Contributors (A-Z)

Prachee Avasthi, Feridun Mert Celebi, Jase Gehring, Megan L. Hochstrasser, Austin H. Patton, Kira E. Poskanzer, Dennis A. Sun, Ryan York

Version 2 · Mar 31, 2025

Purpose

Gene expression drives the identity, behavior, and function of all cells. By comparing gene expression across different species, we can identify genetic networks that are shared or differ across species, allowing us to form hypotheses about the evolutionary origins of diverse cell types. To do this, we must first group similar genes so we can make accurate comparisons. This is traditionally done based on sequence homology. We thought protein structural similarity might provide an alternative, and possibly more relevant, basis for comparing cell type across species.

We explored the performance of gene clusters inferred using either sequence or structural similarity in mixing data from different species and integrating single-cell RNA-seq data from mouse, frog, and zebrafish brain samples. Methods that are able to accurately identify shared genes across species should allow us to identify cell lineages that have shared ancestry — we would expect, for example, that frog, mouse, and zebrafish neurons should express some overlapping set of genes. We found that protein structural clusters preserved data set structure, but these initial attempts did not merge homologous cell types across species better than methods based on sequence homology. While this work was in progress, a conceptually similar approach has apparently succeeded in merging related cell types across species, and we suggest readers familiarize themselves with the protein language model-based method called SATURN **[1]**.

We are no longer actively pursuing this project, but the ideas may be of broad interest, so we are sharing our concept and preliminary results. At the end of the pub, we further discuss potential challenges and opportunities for anyone who may pursue this idea.

- All **code**, including analysis notebooks and outputs, is available in this GitHub repository.
- **Orthogroup and structural cluster files**, plus **feature count matrices** for each species and data set are available on Zenodo.

We've put this effort on ice!

#StrategicMisalignment

We decided not to pursue cross-species single-cell RNA-sequencing analyses in the near term. Since pausing this work, more sophisticated methods have also been released. We may revisit this topic in the future and are excited to see the continued advancements in this field.

Learn more about the Icebox and the different reasons we ice projects.

The motivation

Cells are a fundamental unit of biological organization, so the evolution of cellular function is central to biological research. The conservation and divergence in cellular phenotypes can reveal evolutionary origins and core functional features of cell types, as well as unique innovations displayed in just a subset of species. For example, understanding the evolutionary origins of different cell types in the nervous system may allow us to better understand their physiological function and species-specific differences. What neural genes and pathways are conserved across evolution and which are specific to just one clade? How many neural cell types are shared across species? How much can model nervous systems tell us about the human brain? The answers to these questions may lie in existing gene expression data, but each species is made up of a unique set of genes, preventing direct comparison. A method that places cells from different species in a shared space, in the form of a shared multi-species gene expression matrix, is needed to answer these questions. Such a method would merge cell types across species and unlock cross-species transcriptomics, but this remains a major challenge in bioinformatics **[2]**.

Single-cell RNA-seq atlases — large data sets of single-cell transcriptomes spanning an entire organism — are uniquely suited to study the evolution of cellular function because they offer gene functional information, which we can compare directly across species. These data sets allow us to evaluate how well gene abundance is correlated with annotated cell types. However, comparing gene expression across species requires a shared, multi-species reference space in which we can directly compare gene expression. Genes from diverse species must be grouped into sets of similar genes, which we call **shared feature sets**. The primary gene sequences are different between species, along with overall genome architecture, so there is no straightforward way to merge reference transcriptomes and generate a shared feature set. Many factors, including the list of species being investigated, arbitrary similarity cutoffs, and specific algorithm design choices add complexity to the problem, and as a result, a number of manual and algorithmic methods have been developed and applied to this problem **[3][4][5]**.

Existing methods have often relied on single-copy orthologs — genes with only a single gene copy found in each species used for comparison — making it difficult to compare cell identities when genes have duplicated across different lineages. Moreover, sequence-based methods can fail to detect remote, but shared, ancestry **[6]**. Methods

that avoid relying on single-copy genes or are able to group genes based on expected function, rather than strictly by ancestry, could theoretically improve gene expression comparison across species.

The idea

We decided to explore using protein structural predictions from the AlphaFold Database rather than RNA sequences to create shared feature sets spanning multiple species. We hypothesized that protein structural similarity might outperform gene sequence orthology at merging cell types across species. If protein structure drives protein function more than sequence, then protein structure similarity might better capture functional conservation than sequence similarity across evolutionary distances where remote homology detection is more prone to failure, especially at the cell type level. First, previous approaches are often based on one-to-one-orthologs, while our method creates collapsed groups of related genes based on structural similarity, which may be a more relevant comparison for merging cell types. Second, recognizing that protein structure space is less diverse than protein sequence space, protein structure predictions might better represent protein function than sequence, an idea supported by the recent success in using protein structure predictions for gene functional annotation **[7][8]**. However, as we detail below, our results to date do not appear to collapse cell types from different species into common clusters better than using sequence-based approaches.

While this work was in progress, a related approach based on large protein language models (PLMs) was shown to effectively merge cells from diverse species, and we encourage readers to read their results. The method, SATURN, encodes protein sequences with a protein language model, and proteins in embedded space can be directly compared and clustered **[1]**. In a similar way, our approach uses AlphaFold-predicted structures to make cross-species comparisons, although our initial explorations did not appear to merge cell types effectively. We will discuss how these approaches compare, possible explanations for the present difference in performance, and why they may be superior to sequence homology for this task.

Methods

See detailed methods below or skip straight to the results.

BioFile handling

To coordinate analysis of data across species, we developed a Python package, “biofile_handling,” which allowed us to programmatically organize files from different species into a common structure. For more details, see the biofile_handling documentation page. This package manages the download of files from remote sources and provides an object-oriented way of interacting with diverse collections of biological data. This package also helped standardize data access across Jupyter notebooks to aid in exploratory analysis. We developed this bespoke package due to the specific cloud-based computing strategy we used at the start of this project. In retrospect, some of the core pipelines for this project might have been better implemented using a workflow management system such as Snakemake or Nextflow. For the purposes of reproducing this study, we have left the “biofile_handling” package in place. You can find thorough details on how to reproduce our analyses in our GitHub repository (DOI: 10.5281/zenodo.8264057).

> All **code**, including analysis notebooks and outputs, is available in this GitHub repository.

Data acquisition

We downloaded publicly available single-cell RNA-seq data and cell type annotation files for each species in this study. For each study, we selected a single sample of adult brain scRNA-seq from three species (*Danio rerio* **[9]**, *Xenopus laevis* **[10]**, and *Mus musculus* **[11]**) for our exploratory analyses.

We selected these studies because of the available data features — genes × cells matrices, annotated cell type matrices, and predicted protein structures in the AlphaFold2 database. These data also came from the same technology (Microwell-

seq) which allowed us to more directly compare data from different species without having to worry about platform-specific effects.

To generate peptide files for downstream analysis, we began by identifying the genome version used for each of the original data sets. We used GRCz10 for the zebrafish data, GRCm38.p3 for the mouse data, and JGI-XENLA9.2 from Xenbase for the frog data. For each data set, we downloaded a FASTA file and GFF file of the gene models for that genome and used Transdecoder (version 5.5.0) to generate cDNA files for the genes. We also used Transdecoder to translate cDNA into peptide files using default settings. For each gene model in our data set, we identified a corresponding UniProtKB ID, if available, using the UniProt ID mapping API.

Organism and reference	Study GEO accession	Genome version	Genome FASTA
Zebrafish (*Danio rerio*) **[9]**	GSE130487	GRCz10	GCF_000002035.5_GRCz10_genomic.
Frog (*Xenopus laevis*) **[10]**	GSE195790	JGI-XENLA9.2	XENLA_9.2_genome.fa.gz
Mouse (*Mus musculus*) **[11]**	GSE108097	GRCm38.p3	GCF_000001635.23_GRCm38.p3_gen

Table 1

Public data sources we used in this study.

Generation of shared feature spaces

To compare gene expression across species, we 1) used a common feature space for reference, and 2) assigned genes from each species to that common feature space.

We began by identifying the genome version used for each of the original data sets. We used GRCz10 for the zebrafish data, GRCm38.p3 for the mouse data, and JGI-XENLA9.2 from Xenbase for the frog data.

To identify orthogroups (groups of genes related by ancestry, abbreviated “OG”), we ran OrthoFinder (version 2.5.4) **[12]** using default settings on the Transdecoder peptide files from all three species. We used the orthogroups representing all species (Orthogroups.tsv), which we then used to map genes to orthogroups. For the joint analysis where we mixed cells from all three species, we removed orthogroups that lacked at least one representative gene from all three species.

To identify structural clusters (groups of genes with similar structures, abbreviated “SC”), we downloaded all AlphaFold-v4 structures annotated with each species’ taxid (Drer: 7955; Mmus: 10090; Xlae: 8355). We then used FoldSeek **[13]** to perform all-versus-all TM-score structural comparison for all proteins from all three species. We clustered the all-by-all comparison matrix using the GreedySet algorithm (“Cluster Mode 0”) to generate structural cluster groups as a shared feature space. Of the clustering options offered by FoldSeek, this method provided the most structural clusters for our shared feature space.

Mapping single-cell RNA-seq count data to shared feature spaces

To generate single-cell gene expression matrices in shared feature spaces, gene counts were transferred from the original gene annotation to the appropriate shared feature. For genes that mapped to the same shared feature, we summed the gene expression values per cell. From here, we used the Scanpy **[14]** single-cell analysis package to process count matrices for downstream analysis. For both gene expression and shared feature set data, we used 40 principal components and 50 nearest neighbors as parameters in the `sc.pp.nearest_neighbors` function. For analysis of individual species, cells and genes were filtered as in a standard single-cell RNA-seq workflow. For details, see the included analysis notebooks.

Joint embedding space generation

To generate a joint embedding in either OG or SC feature space, we tried to select features that were differentially expressed in multiple species. We began by identifying the top differentially expressed (“DE”) features for each cluster in the single-species analyses. For our results in Figure 5, we took the top 200 DE features for each cluster

in single-species OG or SC space and generated a list of features for each species individually. We then took the intersection of those feature lists as our “shared DE features" list, which we used to filter our data and build a joint embedding space [see example in this Google Colab notebook]. We used scanpy’s built-in connector Harmonypy to harmonize the gene expression by species. For Figure 5, Supplemental Figure 1, we varied the number of top DE features per cluster from 100 to 300 in increments of 50, following the same approach with identical Scanpy parameters as used in Figure 5.

Results

> **SHOW ME THE DATA**: Orthogroup and structural cluster files, plus feature count matrices for each species and data set are available on Zenodo.

Sequence and structural feature sets capture species-specific transcriptome patterns

To test whether shared feature sets defined by protein structural similarity can merge multi-species single-cell transcriptomes, we analyzed a multi-species data set generated by a single research group using a common library prep methodology **[9][10][11]**. We collected publicly available single-cell RNA-seq data from whole post-embryonic (adult or juvenile) brains of zebrafish (*Danio rerio*), frog (*Xenopus lavis*), and mouse (*Mus musculus*) samples and developed parallel workflows to process gene annotations into two kinds of shared feature sets (Figure 1):

- **OG — Orthology groups**: For sequence comparisons, we used OrthoFinder **[12]** across the three species **[12]** to generate shared feature sets (orthology groups “OG”) (Figure 2, A).
- **SC — Structural clusters**: For structure comparisons, we performed pairwise alignment of genome-wide predicted protein structures from the AlphaFold2 database **[15][16]** for all three species using FoldSeek **[13]**, producing an all-by-all matrix of protein structure similarity scores. Clustering this matrix yielded sets of

structurally similar proteins (structural clusters “SC”), a structural analog to sequence orthologs.

With comparable shared feature sets in hand, we transferred gene counts from the original single-cell RNA-seq count matrices to the cross-species shared feature sets for analysis (see “Methods”). Overall, we have replaced each organism’s specific reference transcriptome with a new, merged reference based on shared feature sets. The distance between cells from different species can be determined based on differential abundance of shared feature sets. We can directly visualize and compare multiple species in this shared space.

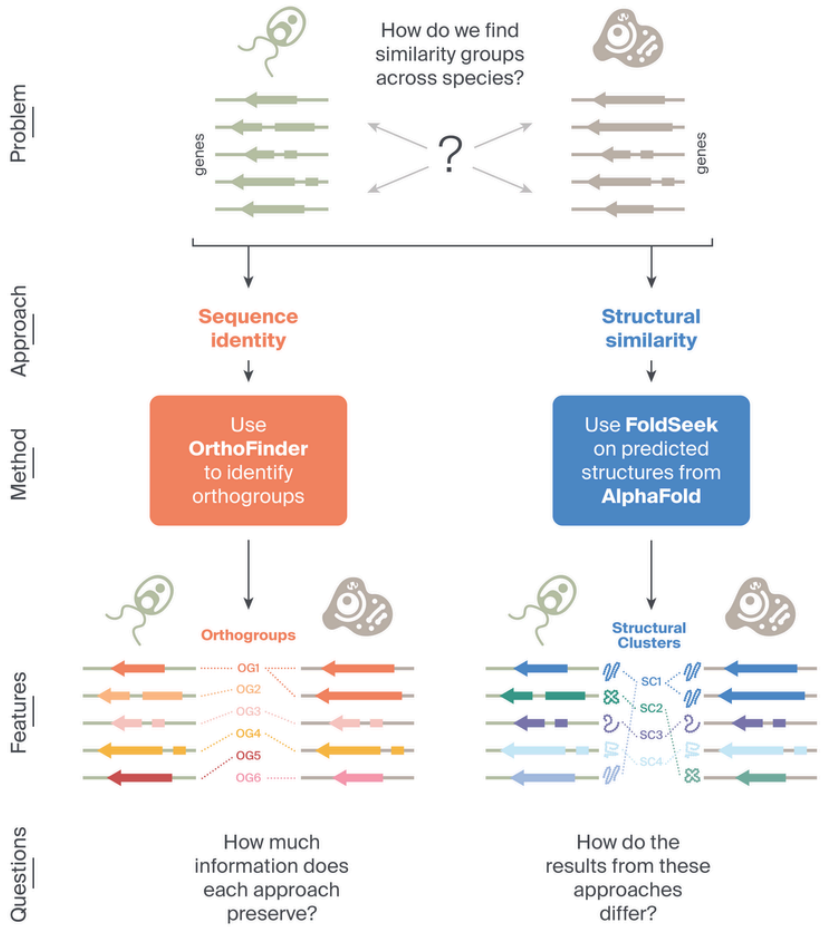

Figure 1

Overall problem and solution strategies.

Comparing genes between species can be difficult, as the composition of each genome varies. How can we identify groups of shared genes between species? We explored two approaches: using OrthoFinder to find groups of genes with similar sequences that are presumably orthologous (diverged from the same ancestral gene copy), or using Foldseek to find groups of genes with similar predicted protein structures. We evaluated how well these groupings preserve biological data and used them to directly compare multiple species.

We first investigated preservation of gene expression features in the orthology group (OG) and structural cluster (SC) feature spaces compared to the original gene

expression space. Compared to gene expression space, OG and SC feature spaces faithfully capture key structures in the data sets, maintaining relationships between cells for a given species (Figure 2, B–J and Figure 3), as judged qualitatively by overall highly similar patterns of clustering and embedding regardless of the shared feature set. Using the un-transformed zebrafish data set as an example, we identified genes using differential expression analysis to determine whether genes appeared to retain their overall expression profile in the new feature spaces. In the original data, a cluster of cells annotated as radial glia and expressing the gene *slc1a2b* is readily apparent (Figure 2, B–C, pink outline). In an embedding of zebrafish cells based on the OG or SC feature sets, the same group of cells, with a similar abundance profile, is marked by representation of orthogroup OG0004873 or structural cluster SC1105 (Figure 2, C, F, I). Similarly, a cluster of cells annotated as neurons, and marked by the gene *aplnra*, is preserved in both OG and SC feature spaces and marked by features specific to the OG and SC spaces, OG0004819 and SC21546 (Figure 2, D, G, J). We also observed a linear relationship between expression of *slc1a2b* and *aplrna* and abundance of their respective OG and SC feature sets (Figure 2, F–G, I–J inset panels). These results suggest that the “expression” of individual genes is preserved in the new feature spaces.

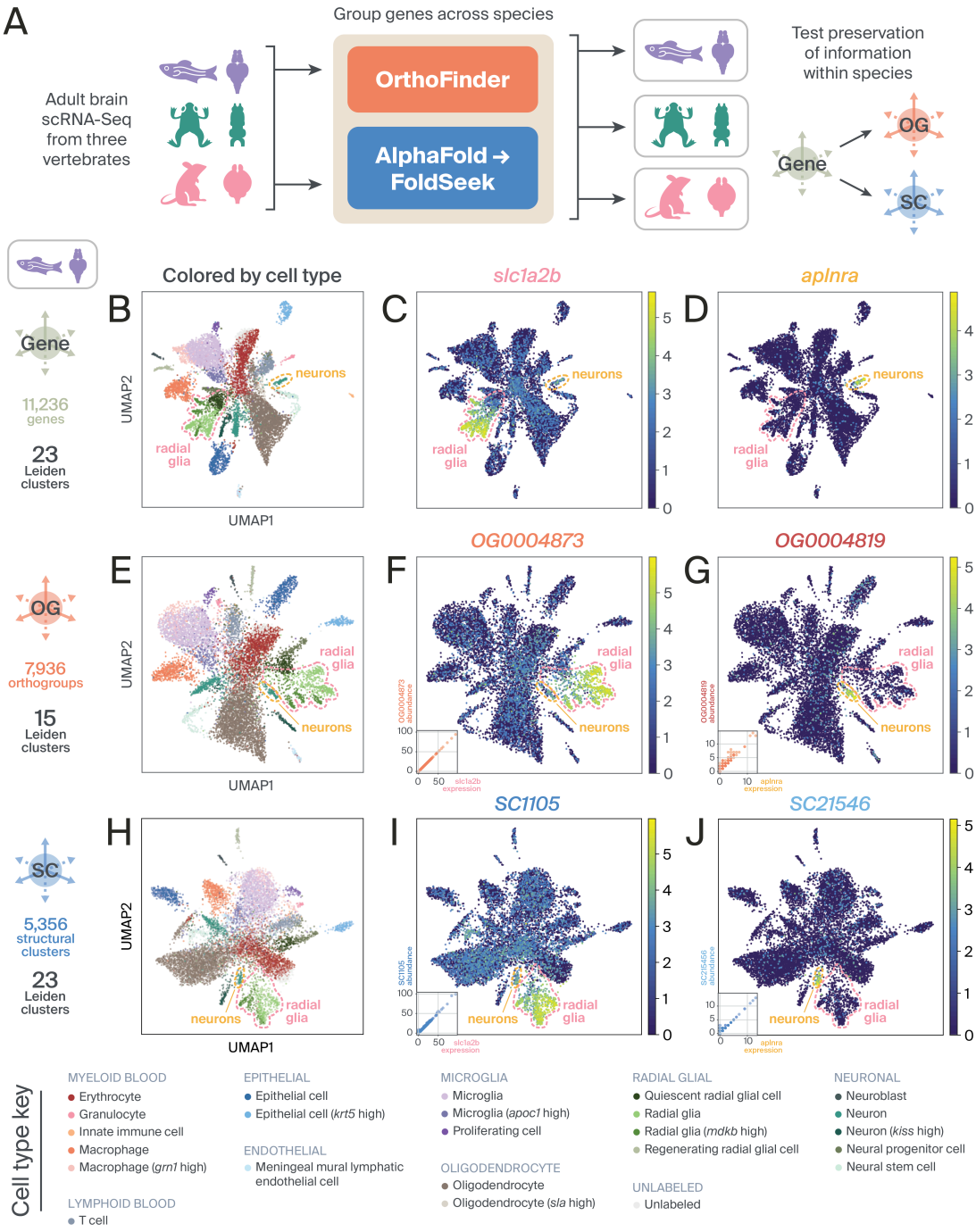

Figure 2

Retention of gene expression information in orthogroup and structural cluster embeddings.

(A) Summary of our overall pipeline for generating shared feature spaces. To generate orthogroup (OG) spaces, we used OrthoFinder. To generate structural cluster (SC) spaces, we clustered AlphaFold structures of proteins using FoldSeek.

(B) UMAP plot of zebrafish cells in gene feature space, colored by cell type.

(C) Plot from (B) colored by expression of *slc1a2b*.

(D) Plot from (B) colored by expression of *aplrna*.

(E) UMAP plot of zebrafish cells in OG feature space, colored by cell type.

(F) Plot from (E) colored by abundance of OG0004873, which contains *slc1a2b*.

(G) Plot from (E) colored by abundance of OG0004819, which contains *aplrna*.

(H) UMAP plot of zebrafish cells in SC feature space, colored by cell type.

(I) Plot from (H) colored by abundance of SC1005, which contains *slc1a2b*.

(J) Plot from (H) colored by abundance of SC21546, which contains *aplrna*.

Inset plots in (F, G, I, J) show correlation in expression of the original gene feature (x-axis) versus the abundance of respective OG or SC feature that contains that gene (y-axis).

–

For information about the contribution of genes from each species to orthogroups and structural clusters, see Supplemental Figure 2.1 (opens in new tab).

Seeking further validation that shared feature sets preserve biological information, we compared cell clusters from our analysis with the published cell type annotations provided by the original authors. These annotations may not comprehensively represent all the cell types present in the data set, but we used them to understand how our embedding spaces could affect interpretation of cell identities. When analyzing cells using the original gene feature space, we recover cell clusters and embedding spaces that are highly similar to the published cell type annotations (Figure 3, A–C). Clustering results are also in broad agreement with the original cell type annotations, meaning that relationships between cells are generally preserved in the

reduced OG and SC spaces (Figure 3, D–I). For example, in the original gene feature space, clusters 1, 4, and 12 predominantly contain microglia, macrophages, and apoc1-high microglia, respectively (Figure 3, C). We observed that when we embedded cells into shared feature spaces, clusters occasionally merged into new clusters. For example, in OG feature space, the three immune cell types above are grouped together in Leiden cluster 1; in SC space (Figure 3, F), these three cell types are grouped in cluster 0 (Figure 3, I). In general, we observed that cell types with known functional similarities (immune cells, glia, neurons) tended to “collapse” together in each of the feature spaces, potentially reflecting shared gene expression signatures between those cell types. Overall, relationships among cells appear to be broadly conserved in our OG and SG shared feature sets.

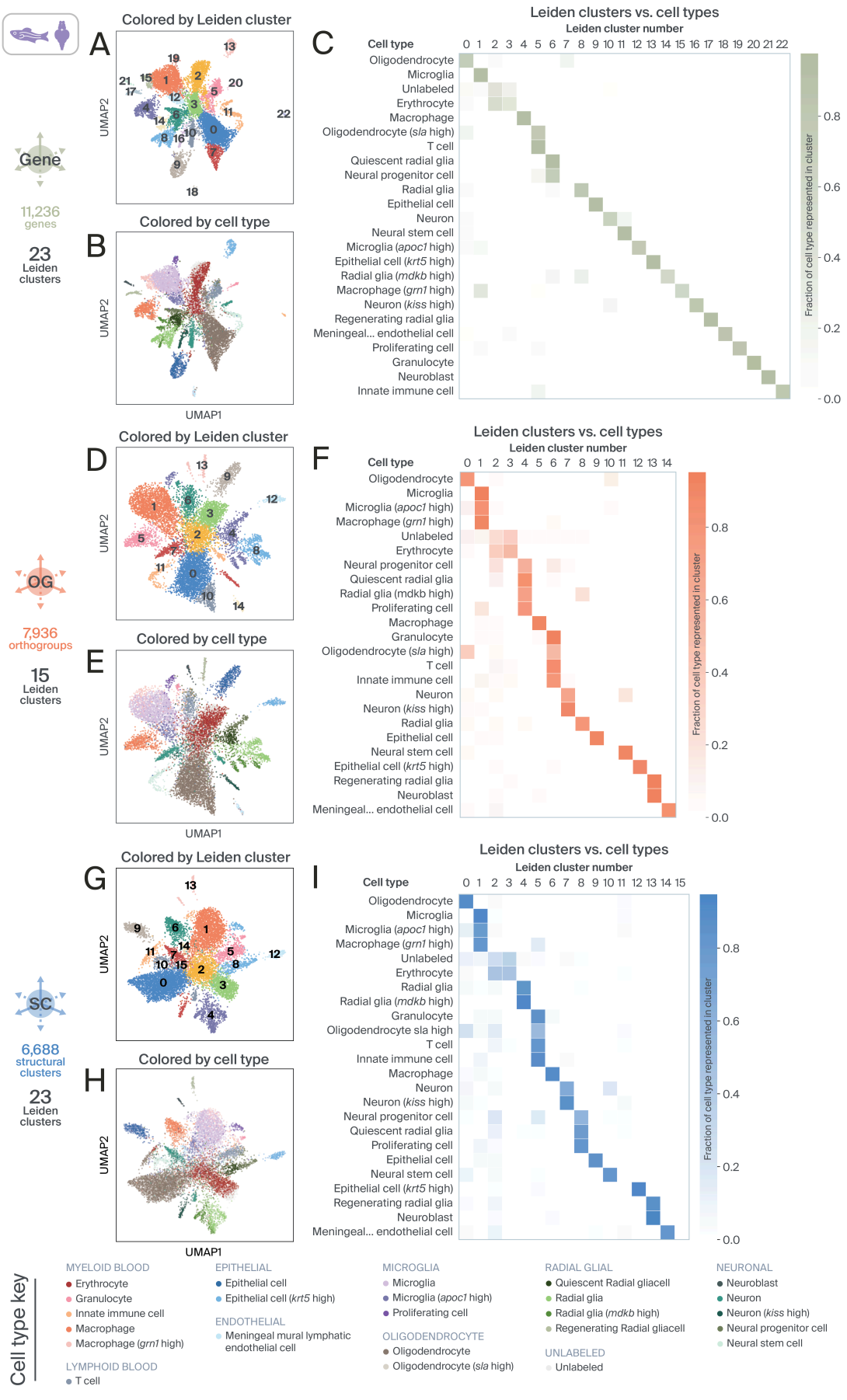
Gene
11,236 genes
23 Leiden clusters
OG
7,936 orthogroups
15 Leiden clusters
SC
6,688 structural clusters
23 Leiden clusters
A
Colored by Leiden cluster
B
Colored by cell type
C
Leiden clusters vs. cell types
Leiden cluster number
Cell type
Fraction of cell type represented in cluster
D
Colored by Leiden cluster
E
Colored by cell type
F
Leiden clusters vs. cell types
Leiden cluster number
Cell type
Fraction of cell type represented in cluster
G
Colored by Leiden cluster
H
Colored by cell type
I
Leiden clusters vs. cell types
Leiden cluster number
Cell type
Fraction of cell type represented in cluster
UMAP1
UMAP2
Cell type key
MYELOID BLOOD
Erythrocyte
Granulocyte
Innate immune cell
Macrophage
Macrophage (grn1 high)
LYMPHOID BLOOD
T cell
EPITHELIAL
Epithelial cell
Epithelial cell (krt5 high)
ENDOTHELIAL
Meningeal mural lymphatic endothelial cell
MICROGLIA
Microglia
Microglia (apoc1 high)
Proliferating cell
OLIGODENDROCYTE
Oligodendrocyte
Oligodendrocyte (sla high)
RADIAL GLIAL
Quiescent Radial gliacell
Radial glia
Radial glia (mdkb high)
Regenerating Radial gliacell
UNLABELED
Unlabeled
NEURONAL
Neuroblast
Neuron
Neuron (kiss high)
Neural progenitor cell
Neural stem cell

Figure 3

Orthogroup and structural cluster embeddings retain cell type information.

(A) UMAP plot of zebrafish cells in gene feature space, colored by Leiden cluster.

(B) Plot from (A), colored by cell type.

(C) Confusion matrix comparing the proportion of cells of each annotated cell type in each Leiden cluster in gene space. Heatmap hue corresponds to the fraction of cells of each cell type that contributed to each Leiden cluster. Each row adds up to 1

(D) UMAP plot of zebrafish cells in OG feature space, colored by Leiden cluster.

(E) Plot from (D), colored by cell type.

(F) Confusion matrix comparing the proportion of cells of each annotated cell type in each Leiden cluster in OG space.

(G) UMAP plot of zebrafish cells in SC feature space, colored by Leiden cluster.

(H) Plot from (G), colored by cell type.

(I) Confusion matrix comparing the proportion of cells of each annotated cell type in each Leiden cluster in SC space.

–

You can browse the heatmaps in this plot interactively by opening these links (opens a new tab):

(C) *Danio rerio* Leiden clusters-vs-celltypes confusion matrix

(F) *Danio rerio* Leiden clusters-vs-orthogroups confusion matrix

(I) *Danio rerio* Leiden clusters-vs-structural clusters confusion matrix

–

For versions of these plots that examine the *Mus musculus* and *Xenopus laevis* data, see Supplemental Figure 3.1 and Supplemental Figure 3.2 (links open in new tabs).

How do cell clusters in OG and SC feature spaces compare to gene space? Are the collapsed clusters functionally meaningful? We used Sankey plots to highlight how clusters of cells are maintained or altered based on the shared feature set used (Figure 4). These plots show the proportion of cells from each original cluster in gene feature space that were placed into each cluster in the shared feature spaces. For each original cluster, we assigned a “primary” destination cluster for which the greatest number of cells in the original cluster arrived in the new feature space, labeled as a gold Sankey plot band. Other destination colors are labeled using a beige Sankey plot band.

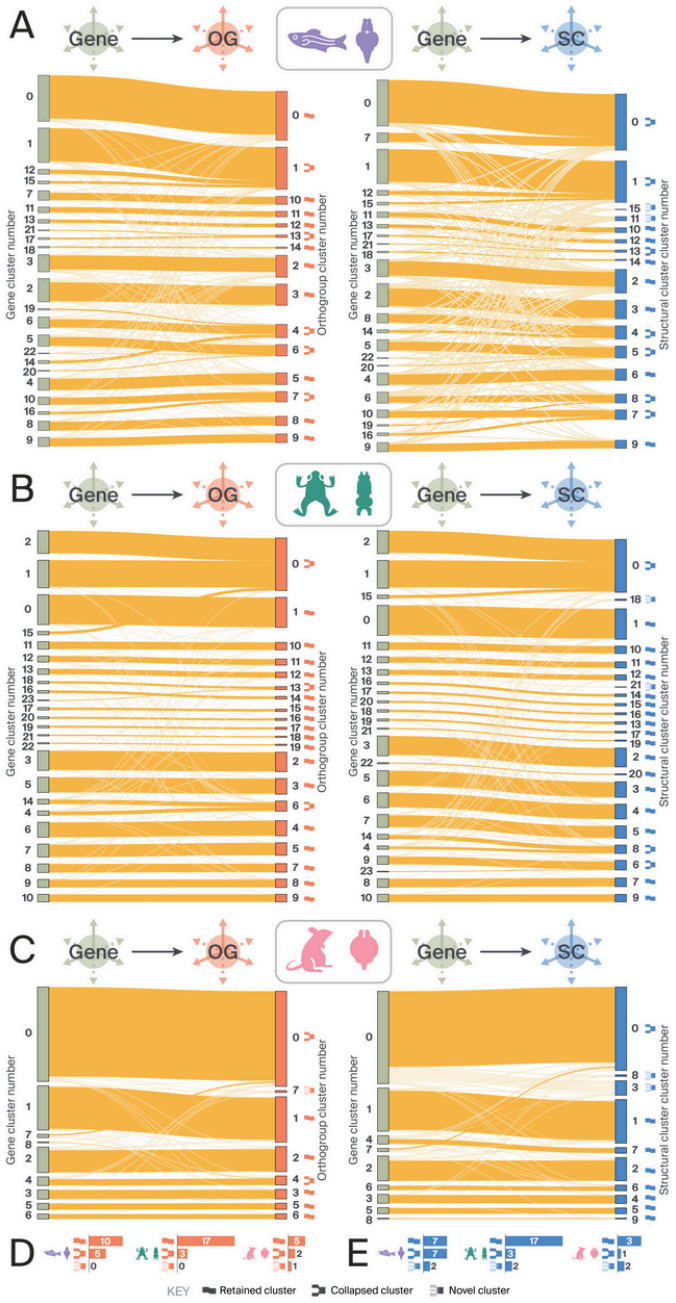

Figure 4

Cell clustering effects in orthogroup and structural cluster embeddings differ by species.

For each row, the left plot shows the comparison of Leiden clusters in gene space to Leiden clusters in OG space.

The right plot shows the comparison of Leiden clusters in gene space to Leiden clusters in SC space.

To the right of each plot, shared feature space clusters are annotated with one of three icons indicating that it is a “retained” cluster, a “collapsed” cluster, or a “novel” cluster – see KEY at the bottom of figure.

(A) Sankey plots for zebrafish cells.

(B) Sankey plots for frog cells.

(C) Sankey plots for mouse cells.

(D) Summary of proportion of retained, collapsed, and novel clusters for each species in OG space.

(E) Summary of proportion of retained, collapsed, and novel clusters for each species in SC space.

–

You can browse the Sankey plots in this image interactively by opening the following links (opens in a new tab):

(A, left) *Danio rerio* genes-vs-orthogroups Sankey plot

(A, right) *Danio rerio* genes-vs-structural clusters Sankey plot

(B, left) *Mus musculus* genes-vs-orthogroups Sankey plot

(B, right) *Mus musculus* genes-vs-structural clusters Sankey plot

(C, left) *Xenopus laevis* genes-vs-orthogroups Sankey plot

(C, right) *Xenopus laevis* genes-vs-structural clusters Sankey plot

We categorized each cluster based on whether it was 1) a “retained” cluster with cells primarily from a single cluster in gene space; 2) a “collapsed” cluster with major contribution of cells from multiple clusters in gene space; or 3) a “novel” cluster with minor contributions of cells from multiple clusters in gene space (Figure 4). Notably, while we observed many collapsed clusters in both OG and SC spaces in all three species, we observed an enrichment of novel clusters containing mixtures of cells from multiple cell types in the SC space. (Figure 4, D–E). The relative proportion of collapsed versus novel clusters varied between species. For example, zebrafish cells in OG feature space produced five collapsed clusters and zero novel clusters, whereas the same cells in SC feature space produced nine collapsed clusters and eight novel clusters. Gene, OG and SC feature spaces are not equivalent, and collapsing based on feature similarity is a potentially valuable way to embed and understand cellular information.

Overall, we observed that embedding cells into different shared feature spaces resulted in broadly concordant patterns of clustering and preservation of feature abundance. By converting genes to either orthogroups or structural clusters, we introduced some degree of distortion to our data in a signal-dependent manner. The degree and nature of this distortion varied in the orthogroup and structural cluster feature spaces, and we have not systematically explored OG and SC clustering parameters to understand these spaces well. Shared feature abundance can often be rationalized in terms of gene expression, and each species appears to embed well into its species-specific OG or SC feature space. We next turned to embedding multiple species simultaneously.

Embedding multiple species only marginally merges cell types

After confirming that orthogroups (OG) and structural clusters (SC) largely preserve structure in scRNA-seq data, we investigated how these shared feature sets might facilitate cross-species analysis of cell identity and feature set abundance. We implemented a pipeline to generate a feature set capable of mixing cells from different species using OG or SC feature spaces into a joint embedding (Figure 5, A). To create a list of features used in the joint embedding, we began by identifying the top 200 most differentially expressed (DE, see “Methods”) features in the single-species OG or SC analyses. We took the lists of top DE features from each species and used the

intersection of features between all three species — the most differentially abundant and mutually shared features — and used this list as the starting point for our analyses. We thought that selecting only features which are differentially expressed in all species would reduce differences between species and produce a more merged embedding space. After filtering out low-abundance features and identifying highly variable genes through Scanpy (see "Methods"), we used the Harmony Python package to "batch-correct" features by species. This resulted in a joint embedding space where we could jointly examine cells from multiple species.

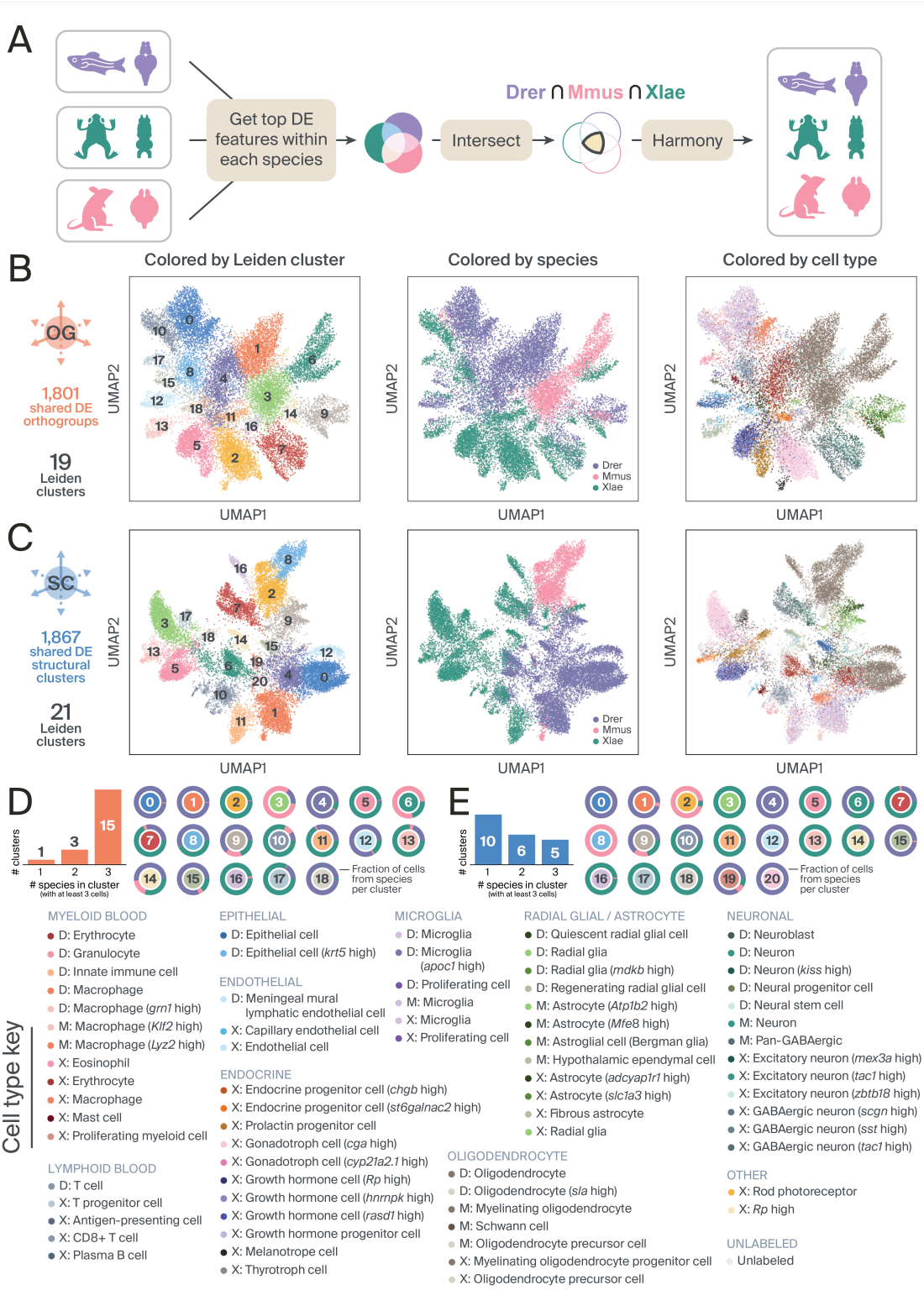

Figure 5

A pipeline using the intersection of within-species differentially expressed genes and the Harmony algorithm allows for joint embedding of cells across species.

Drer: *Danio rerio*, Mmus: *Mus musculus*, Xlae: *Xenopus laevis*

(A) Summary of pipeline for generating joint cell embeddings across species.

(B) UMAP plots of cells from all three species in OG feature space, colored by Leiden cluster, species, and cell type respectively.

(C) UMAP plots of cells from all three species in SC feature space, colored as in (B).

(D) Summary of species composition for Leiden clusters in OG space. Ring plots on the right show the proportion of cells from each species per Leiden cluster. Bar plot on the left shows the number of clusters with 1, 2, or 3 species' cells present.

(E) Summary of species composition for Leiden clusters in SC space, as in (D).

–

For a breakdown of how the number of "top genes" used for generating the joint embedding affects the mixing of cells, see Supplemental Figure 5.1 (opens in new tab).

You can browse the scatter plots from Figure 5 interactively using the widgets below:

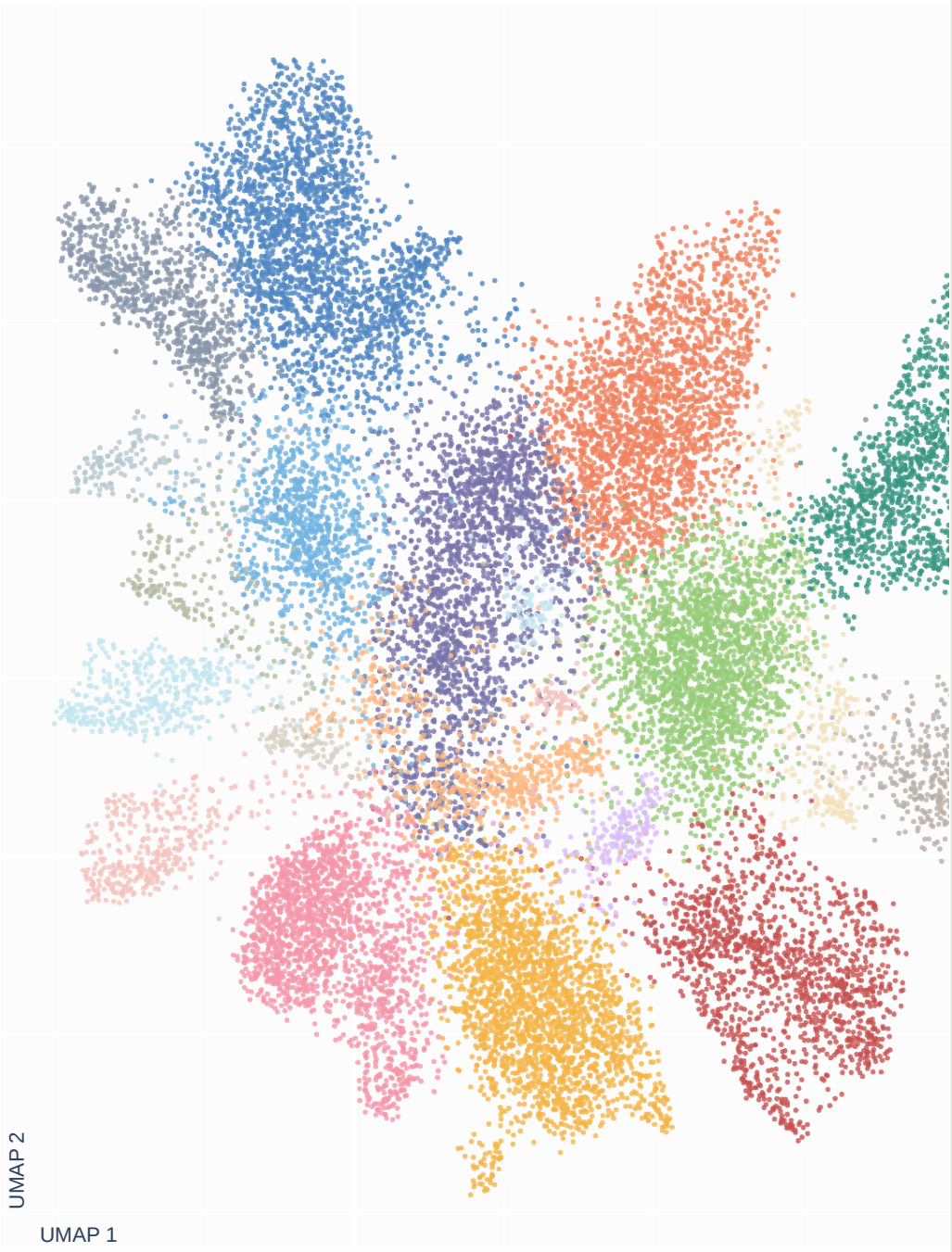

Figure 6

Joint embedding of cells from mouse, frog, and zebrafish brains in orthogroup space.

Drer: *Danio rerio*, Mmus: *Mus musculus*, Xlae: *Xenopus laevis*

Each cell in this space is represented as a point. You can hover over the points to see the cell barcode, cell type, and Leiden cluster associated with each cell. You can also toggle the data overlay that colors the plot using the drop-down menu to switch between views that color cells by Leiden cluster, species, and cell type. Clicking and selecting an area allows you to zoom in on a group of cells. Double-clicking returns the zoom to the original size. Clicking on an entry in the legend below the drop-down menu toggles the visibility of each group of cells. Double-clicking on an entry hides all categories other than the selected group.

Figure 7

Joint embedding of cells from mouse, frog, and zebrafish brains in structural cluster space.

This plot shows the same cells from the previous interactive plot based on their position in the structural cluster embedding. Notably, cells from the three species are not as well-mixed as in the orthogroup space.

We compared and analyzed the resulting OG and SC joint embedding spaces to understand their relative performance (Figure 5, B–C). In OG space, cells from three species appeared to be moderately mixed, with analogous cell types from different species appearing to mix in the same clusters (Figure 5, B). SC space showed some, but noticeably less, mixing across species. For example, radial glial and astrocyte cell identities from all three species appeared to be mixed in OG cluster 9. Among the 19 clusters in the OG space, 15/19 (79%) contained cells from all three species in the analysis (Figure 5, D), compared to only 5/21 (24%) in the SC space (Figure 5, E).

The apparent difference in performance between OG and SC space in creating "mixed" clusters could be caused by a variety of factors, including the representation of different species in each shared feature group, the size distribution of feature groups, and many other parameters. To understand the impact of the starting shared DE feature count on each embedding space, we sampled the top 100, 150, 200, 250, or 300 genes from each cluster from each species in either OG or SC space, and used these lists as starting points for new embedding spaces (Figure 5, Supplemental Figure 1). We observed that varying the number of top DE features we used had a linear relationship with the resulting number of starting DE features for the shared feature space. The number of shared DE features in both OG and SC space remained comparable (Figure 5, Supplemental Figure 1, A, C). Notably, OG spaces appeared to generally have more clusters containing cells from all three species, whereas SC spaces appeared to have more clusters containing cells from just one species across all our analyses. This suggests that the differences in OG and SC performance are robust to variations in the starting number of shared DE features used to build the joint embedding.

To further examine the harmonization of cells across species in the joint embedding spaces, we examined the degree of concordance in cell type annotations across the three species (Figure 6). We observed that in the OG joint embedding space, a few clusters displayed similar representation across all three species. For example, OG cluster 0 contained immune cells (macrophages and microglia) from all three species; OG cluster 7 contained neurons from all three species; and OG cluster 9 contained radial glia and astrocytes from all three species. These results suggest that our OG

joint embedding pipeline appears to be able to mix cells of different species to some degree. We saw comparable clusters in SC space for the three broad cell types found in OG space: SC cluster 1 contained immune cells, SC cluster 7 contained neurons, and SC cluster 9 contained radial glia and astrocytes. These results suggest that while OG and SC spaces seem to differ in their ability to mix cells across many identities, there may be “core” groups of features that are readily comparable between the two approaches.

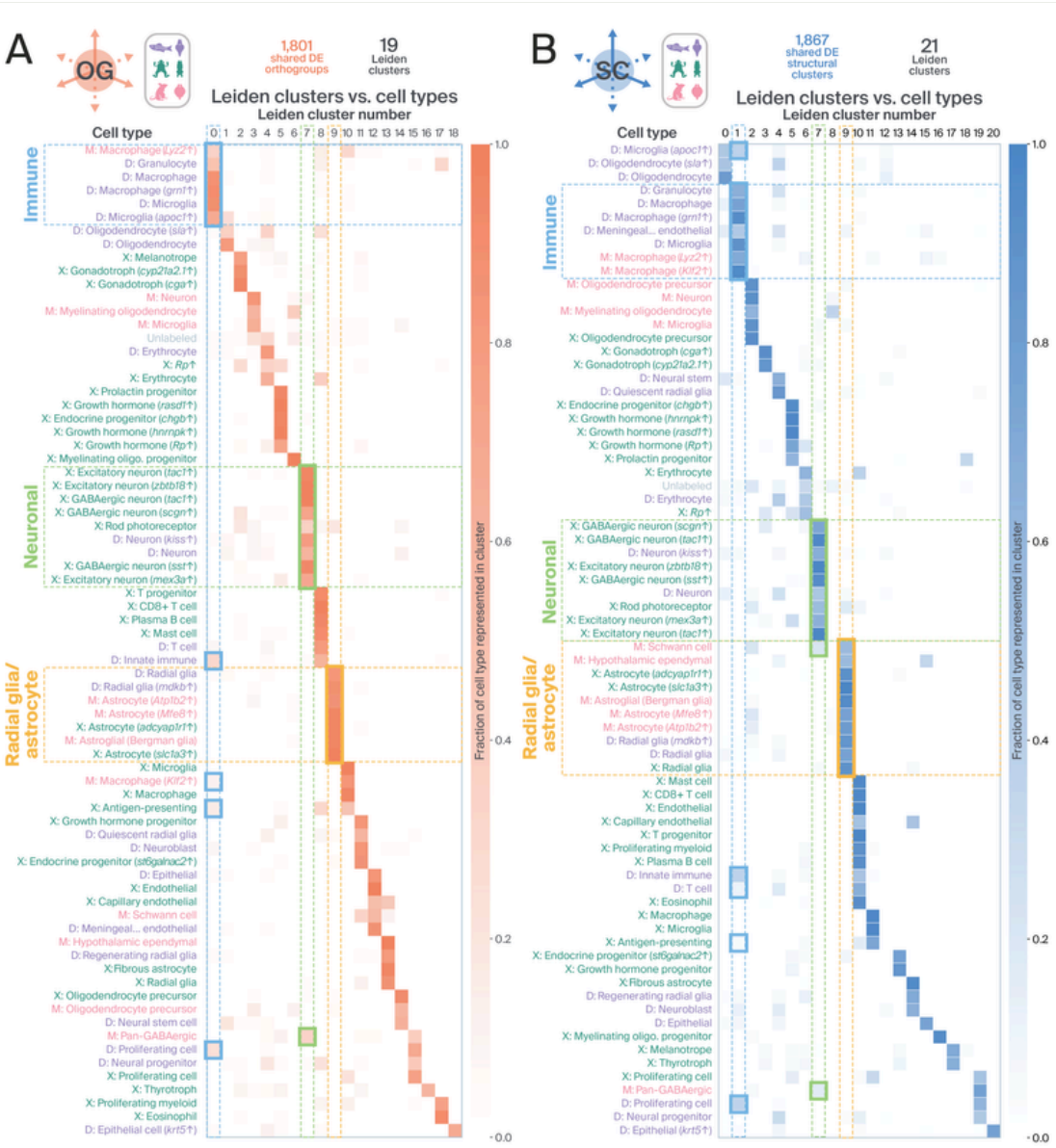

Figure 8

Multiple cell identities are shared between species using both sequence identity and structural similarity approaches.

(A) Confusion matrix comparing the proportion of cells of each annotated cell type from all three species in each Leiden cluster in OG space. Heatmap hue corresponds to the fraction of cells of each cell type that contributed to each Leiden cluster. Each row adds up to one. Dotted lines highlight three clusters of interest that appear to be composed of Immune, Neuronal, and Radial glial/ astrocyte identities. Solid boxes highlight cells that appear to contribute to the identity of the cluster.

(B) Confusion matrix comparing the proportion of cells of each annotated cell type from all three species in each Leiden cluster in SC space, as in (A). Dotted lines highlight three clusters of interest that appear to be composed of immune, neuronal, and radial glial/astrocyte identities. Solid boxes highlight cells that appear to contribute to the identity of the cluster.

–

You can browse the heatmaps in this plot interactively through the following links (opens a new tab):

(A) Confusion matrix of cell annotations from mouse, zebrafish, and frog versus Leiden clusters in orthogroup space.

(B) Confusion matrix of cell annotations from mouse, zebrafish, and frog versus Leiden clusters in structural cluster space.

Discussion

We have developed an approach to merge cell types across species using protein structural similarity as a basis for comparing gene expression. By mapping genes to groups of similar protein structures, we cast transcriptomes into a common reference space, even when starting with multiple species. Our initial investigations presented here only partially merge cells from multiple species from a Microwell-seq data set. There are many potential reasons for the overall lack of merging, and we discuss some of these in the "Challenges" section below.

Shared feature set performance

Shared feature sets have two purposes in the context of this study. First, they are interesting objects in which to compare genes across and within species. Depending on the method of compression, we may be able to infer evolutionary relationships like orthology, convergence, and functional duplication. Second, shared feature sets are useful for mapping cell types across species from single-cell gene expression data. We have shown that mapping reference transcriptomes to groups of orthologous (OG)

or similar (SC) genes preserves single-cell transcriptional information, and it is very natural to produce OG and SC shared feature sets that span multiple species. In the case of SATURN, protein embeddings from a protein language model are clustered into yet another type of shared feature space, with apparently very good results in merging cells across species.

But there are limits to the utility of shared feature spaces. They can distort gene expression in complex ways, and they do not represent true functional equivalence across or within species. It is important to note that shared feature sets (or "macrogenes" in the parlance of SATURN) do not reduce the burden of exploring single-cell expression data at the gene level, and including multiple species only increases the complexity of these data sets. However, shared feature sets are useful for identifying homologous cell types, and they conveniently group genes in a meaningful way that is useful for downstream gene-level analysis.

Our efforts thus far yield some degree of mixing between cells of different species in joint embedding spaces. While it is possible to combine cells from many species into one data set, clustering and embedding algorithms easily identify species-specific differences. Our attempts to use regression or a variety of single-cell integration methods generally fail to merge cell types across species. Feature set selection and batch correction to force merging between species may produce misleading results.

While it is important to merge homologous cell types, we believe it will be more useful for the community to explore methods that compare cell types with minimal distortion to the underlying gene expression data. Once homologous cell types have been identified, we need downstream tools that are able to make cross-species comparisons on data that has *not been batch-corrected*. After all, these cells are from different species and the differences that are regressed out for the purposes of merging could be biologically meaningful. Rather, we need to understand the components of gene expression that drive differences across species — adaptation, drift, technical effects, functional compensation, etc. — to reconstruct the history and meaning of cell function evolution.

Further work is required to determine if the OG and SC feature spaces are in fact very good spaces in which to compare multiple species. Unfortunately, we lacked a working example of merging across species during the development of this work, so it was difficult for us to debug our approach. If we were to resume this project, we would start by attempting to reproduce the results of the SATURN paper, followed by evaluating

the differences between shared feature sets based on sequence orthogroups, protein language models, and protein structural models in a more controlled setting.

It is also possible that, even with further development, we may discover fundamental differences in the nature of structure- and sequence-based comparisons. One interpretation of our results could be that the absence of cell type merging using structures is caused by fundamental biological differences. Given that sequence–structure relationships are known to be nonlinear, structures might actually be more dissimilar than expected based on sequence (e.g., **[17][18]**). Therefore, failure to merge cell types using structures could be indicative of true functional differences in cell behavior or physiology. Across evolutionary time, the relationship between cell identity and structure might differ. Such possibilities are ripe for future exploration.

Merging cell types across species

When comparing cells across species in a shared reference space, a perfect merge or overlap of cell types across species is not necessarily desirable. Biological differences between species should be preserved, and it may be expected that homologous cell types will not merge together in a shared embedding space. Furthermore, efforts to force data into constrained topologies can introduce artifacts and mask real biology. Finally, methods that claim to integrate or harmonize data from multiple experiments cannot distinguish between biological and technical effects, and they must be employed with caution in the course of single-cell analysis. Rather than methods that can mash cells into recognizable clusters, we need high-quality data sets that can be compared with minimal batch correction or distortion, along with workflows that recognize when to employ batch-corrected versus uncorrected count data.

Comparison to SATURN

The SATURN package introduces a concept of “macrogenes” that are exactly analogous to the shared feature sets discussed here. Rather than grouping genes according to their corresponding protein structural similarities, SATURN instead creates macrogenes based on protein embedding similarity. Protein embeddings are the output of protein language models. They are a vector representation of the protein sequence, and crucially these vectors can be directly compared in protein embedding

space. We in fact wanted to try this approach in the course of our work, but we focused on protein structures as they are widely available via the AlphaFold database. In addition to using protein language models, SATURN employs sophisticated methods to weight the contribution of each gene to the set of macrogenes, and they employ an autoencoder to generate latent cell embeddings while we used more standard dimensionality reduction and batch-correction methods. We have not yet deeply examined the performance of SATURN or been able to compare the performance of protein embeddings versus protein structural predictions in this setting. We are very encouraged by the results from SATURN, and we look forward to exploring its capabilities and putting it to use.

Challenges

We faced a number of challenges in the course of working on this project, many of which were technical rather than scientific.

1. **Data acquisition and sanitization**. For the analysis we've shared in this pub, we used data generated primarily by a single laboratory (the Guo lab at Zhejiang University) on a common sequencing platform (Microwell-seq). However, during the course of this work, we also downloaded and explored data from many different sequencing platforms (Drop-seq, 10x Chromium, inDrop), organisms (mouse, human, bearded dragon, turtle, frog, axolotl, salamander, zebrafish), and research groups. We observed that sequencing data from different research groups had considerable variability in the availability of code, accessibility of data, quality of documentation, and formatting of files. These differences make it challenging to reproduce or even understand previously published work.
2. **Data quality**. Among the data sets we examined, there was also substantial variability in data quality. Some data sets contained large numbers of low-read count cells, or large numbers of small samples that required batch correction. Without analyzing the data, it was not usually straightforward to know whether it would be useful. The difficulty in accessing data from different sources was an additional barrier to analysis.
3. **Computational infrastructure**. Single-cell sequencing produces very large files (tens to hundreds of GB) that require large amounts of RAM to load and analyze. We ultimately used the Cloud9 platform from Amazon Web Services to generate

remote computing environments capable of analyzing these data, but analyzing moderately-sized scRNA-seq data (~10,000 cells) requires moderately powerful computing (> 16 GB RAM), which can be a barrier to exploring this type of data.

Overall, these challenges brought to light a contradiction in the current state of single-cell sequencing studies. Many papers argue that the utility of their work comes from generating data resources for the broader scientific community. Yet these same studies often provide few practical ways to access and analyze their data. Others have highlighted this contradiction through a variety of meta-analyses **[19][20]**.

From our experiences working with these data, we believe that single-cell RNA sequencing studies could benefit from the following changes to make the data more usable:

1. **Standardization of data formats**. Read count matrices for scRNA-seq data are archived in formats including plain text TXT, CSV, or TSV files; platform-specific CellRanger, or Drop-seq formats; compressed formats such as LOOM or H5AD; and numerous other schema. To facilitate ease of data access, we would recommend scientists share read count matrices as CSV files or H5AD files, as these formats are more broadly accessible to those intending to utilize the data. Often forgotten, gene names must be included with the genes × cells matrix, as a separate file or as the index of a data matrix.
2. **Improved software and code documentation**. Software and code used for single-cell sequencing analysis vary in their level of documentation. This variability can be a consequence of the degree of familiarity of individual researchers with programming, the amount of time that documentation requires, and limited oversight of code quality and documentation. We would recommend that authors use GitHub to centralize code for their analyses and Conda, Docker, or executable notebooks (Binder, Google Colab) to manage computing environments. Free and publicly available resources through The Carpentries and other organizations can help researchers less experienced with programming to make their code and software more accessible. For all projects, downstream users should be able to reproduce and extend the initial analysis, requiring planning for new users, posting to public repositories, and providing documentation, code, and working examples or tutorials.
3. **Greater oversight**. If cell atlases and other scRNA-seq studies are to live up to their oft-promised impact, it is imperative that researchers and publishing

organizations hold each other accountable for producing work that is useful to a broad variety of scientists. Some publishers have adopted frameworks for data sharing such as the eLife MDAR Framework. Data and metadata standards catalogs such as FAIRsharing could help produce better guidelines for data access and reproducibility for scRNA-seq data.

We've tried to provide relatively comprehensive data, code, and software documentation following the goals laid out by Arcadia's Software team:

1. **Data**. You can access the orthogroup and structural cluster files, as well as the feature count matrices for each species and data set, in this Zenodo record.
2. **Code**. The code used for generating all of our analyses and components of figures is available in this GitHub repository.
3. **Software documentation**. In our GitHub repository, we explain how we conducted our analysis through a collection of Jupyter notebooks. The software packages we used for our analysis are collected in this Conda environment.
4. **Data exploration**. We have built a number of interactive HTML visualizations for our data using Plotly embedded in the text of this pub, as well as a Google Colab notebook that users can use to download and explore the joint embedding spaces generated by this analysis.

Next steps

We decided to "ice," or pause, this project in the course of our changing research priorities. Our code and documentation serve as a snapshot of this project and contain areas that are incomplete or suboptimal, such as the "biofile_handling" framework. However, we can always pick up our work in the future as the need arises.

We plan to apply the techniques and insights we gained from this exploration to more near-term efforts within Arcadia. For example, we're interested in continuing to use comparisons between protein sequence and structure embeddings to understand the function of diverse genes **[21]**. The challenges we faced in using publicly available code and data have also strengthened our commitment to making our software usable and reproducible.

We're sharing these preliminary results as part of our commitment to open science and to maximizing the utility of our work. While we would need continued work to fully evaluate and understand these methods, we hope the analyses and code we've shared can be a starting point for others interested in exploring this space.

Opportunities for follow-up

If you are interested in building on this foundation, we suggest more quantitatively comparing this approach to standards in the field such as SAMap or SATURN, exploring the parameter space of clustering algorithms, and testing these methods on diverse data sets.

Weigh in!

We'd love to hear from you, especially about the following: Have you used structural homology as an alternative to sequence homology in your research? Do existing cell atlases contain sufficient depth of coverage in cell types and transcriptomes to make evolutionary comparisons? How can our community improve its data collection and sharing practices to make meta-analyses like this more tractable?

If you have thoughts to share, please don't hesitate to leave a comment!

References

1. Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, Leskovec J. (2023). Towards Universal Cell Embeddings: Integrating Single-cell RNA-seq Datasets across Species with SATURN. https://doi.org/10.1101/2023.02.03.526939
2. Shafer MER. (2019). Cross-Species Analysis of Single-Cell Transcriptomic Data. https://doi.org/10.3389/fcell.2019.00175

3 Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, Wang B. (2021). Mapping single-cell atlases throughout Metazoa unravels cell type evolution. https://doi.org/10.7554/elife.66747

4 Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G. (2018). Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. https://doi.org/10.1126/science.aar4237

5 Wang R, Zhang P, Wang J, Ma L, E W, Suo S, Jiang M, Li J, Chen H, Sun H, Fei L, Zhou Z, Zhou Y, Chen Y, Zhang W, Wang X, Mei Y, Sun Z, Yu C, Shao J, Fu Y, Xiao Y, Ye F, Fang Xing, Wu H, Guo Q, Fang Xiunan, Li X, Gao X, Wang D, Xu P-F, Zeng R, Xu G, Zhu L, Wang L, Qu J, Zhang D, Ouyang H, Huang H, Chen M, NG S-C, Liu G-H, Yuan G-C, Guo G, Han X. (2022). Construction of a cross-species cell landscape at single-cell level. https://doi.org/10.1093/nar/gkac633

6 Weisman CM, Murray AW, Eddy SR. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. https://doi.org/10.1371/journal.pbio.3000862

7 Ruperti F, Papadopoulos N, Musser JM, Mirdita M, Steinegger M, Arendt D. (2023). Cross-phyla protein annotation by structural prediction and alignment. https://doi.org/10.1186/s13059-023-02942-9

8 Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. (2021). Structure-based protein function prediction using graph convolutional networks. https://doi.org/10.1038/s41467-021-23303-9

9 Jiang M, Xiao Y, E W, Ma L, Wang J, Chen H, Gao C, Liao Y, Guo Q, Peng J, Han X, Guo G. (2021). Characterization of the Zebrafish Cell Landscape at Single-Cell Resolution. https://doi.org/10.3389/fcell.2021.743421

10 Liao Y, Ma L, Guo Q, E W, Fang Xing, Yang L, Ruan F, Wang J, Zhang P, Sun Z, Chen H, Lin Z, Wang Xueyi, Wang Xinru, Sun H, Fang Xiunan, Zhou Y, Chen M, Shen W, Guo G, Han X. (2022). Cell landscape of larval and adult Xenopus laevis at single-cell resolution. https://doi.org/10.1038/s41467-022-31949-2

11 Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, Huang D, Xu Y, Huang W, Jiang M, Jiang X, Mao J, Chen Y, Lu C, Xie J, Fang Q, Wang Y, Yue R, Li T, Huang H, Orkin SH, Yuan G-C, Chen M, Guo G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. https://doi.org/10.1016/j.cell.2018.02.001

12 Emms DM, Kelly S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. https://doi.org/10.1186/s13059-019-1832-y

13 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2022). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1101/2022.02.07.479398

14 Wolf FA, Angerer P, Theis FJ. (2018). SCANPY: large-scale single-cell gene expression data analysis. https://doi.org/10.1186/s13059-017-1382-0

15 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

16 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

17 Hark Gan H, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, Schlick T. (2002). Analysis of Protein Sequence/Structure Similarity Relationships. https://doi.org/10.1016/s0006-3495(02)75287-9

18 Rahman RS, Rackovsky S. (1995). Protein sequence randomness and sequence/structure correlations. https://doi.org/10.1016/s0006-3495(95)80325-5

19 Deshpande D, Sarkar A, Guo R, Moore A, Darci-Maher N, MANGUL S. (2021). A comprehensive analysis of code and data availability in biomedical research. https://doi.org/10.31219/osf.io/uz7m5

20 McGuinness LA, Sheppard AL. (2021). A descriptive analysis of the data availability statements accompanying medRxiv preprints and a comparison with their published counterparts. https://doi.org/10.1371/journal.pone.0250887

21 Avasthi P, Bigge BM, Celebi FM, Cheveralls K, Gehring J, McGeever E, Mishne G, Radkov A, Sun DA. (2024). ProteinCartography: Comparing proteins with structure-based maps for interactive exploration. https://doi.org/10.57844/ARCADIA-A5A6-1068