Benchmarking strategies for cross-species integration of single-cell RNA sequencing data

Yuyao Song [1, *], Zhichao Miao [1], Alvis Brazma [1], Irene Papatheodorou [1, *]

1 European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, United Kingdom

* Corresponding authors

Emails:

Yuyao Song: ysong@ebi.ac.uk

Irene Papatheodorou: irenep@ebi.ac.uk

Address:

EMBL's European Bioinformatics Institute (EMBL-EBI)

Wellcome Genome Campus

Hinxton, Cambridgeshire

CB10 1SD

United Kingdom

Tel:

+44(0)1223 494 444

Abstract

The emergence of multiple species cell atlases has opened the opportunity to investigate evolutionary relationships at the cell type level and understand similarities and differences between cell types across species. Cross-species integration of single-cell RNA-sequencing data has been a key approach to comparing cell types between species. To robustly identify homologous cell types and species-specific cell populations, it is crucial to obtain reliable and informative integration. However, currently there are neither benchmarks nor guidelines on how to choose among a diversity of integration strategies. To examine the output characteristics of different cross-species integration strategies, we benchmarked 25 combinations of gene homology mapping methods and integration algorithms in a variety of biological settings. We examine each strategy's capability to perform species-mixing of known homologous cell types and preserve biological heterogeneity using 10 metrics, among which we developed a new biology conservation metric to address the maintenance of cell type distinguishability. Overall, scVI, SeuratV4 methods and SAMap achieve a balance between species-mixing and biology conservation. For evolutionarily distant species, including in-paralogs and using nearest-neighbour based methods is beneficial. SAMap outperforms when integrating whole-body atlases and can prioritize identity-defining similarities between homologous cell types in complex circumstances. We provide our freely available cross-species integration and assessment pipeline to help identify the optimum strategy to analyse new data and develop new algorithms.

Introduction

Animal cells show conspicuous diversity and have fundamental unities across species. Recently, single-cell RNA-sequencing (scRNA-seq) and single-nucleus RNA-sequencing (snRNA-seq) have been applied in a diversity of species, generating full-scale cell atlases from transcriptome profiles of millions of individual cells [1–7]. Comparing cellular expression profiles cross-species provides insights into the origin and evolution of organs and cell types [6,8–10], highlights species-specific transcriptomic innovations [11], as well as further delineates how cell types execute their functions in health and disease [12,13].

Driven by millions of years of evolution, the gene expression profiles of evolutionarily related and functionally similar cell types from different species exhibit significant global transcriptional shifts [11,14,15]. Therefore, cells of different types from the same species usually show higher overall similarity, compared with their cross-species counterparts, in joint analysis of scRNA-seq/snRNA-seq data [16,17]. One key approach to match cells with similar expression profiles from different species is to perform data integration: to find a latent representation of the cells, in which cells are grouped by their transcriptional identity regardless of species of origin. Integrated data enables identification of homologous cell types and rare populations, detection of species-specific cell types, characterization of cross-species differential expression genes in shared cell types, as well as other analyses. Accordingly, an informative integration should achieve mixing of homologous populations between species, separation of species-specific populations, and maintenance of cell type distinguishability on the latent representation.

Various data integration methods originally designed for batch correction or cross-condition integration have been rapidly adapted to cross-species analysis [15]. fastMNN is based on identifying mutual nearest neighbours (MNNs) cross datasets to correct each dimension in principle component (PC) space [18]; harmony performs iterative clustering to impute batch correction vectors based on putative cell type centroids in PC space [19]; LIGER uses integrative non-negative matrix factorization (iNMF) to learn shared and unique gene expression signatures between datasets, while its recent upgrade LIGER UINMF also include unshared features during iNMF to account for species-specific genes [20]; scanorama is a panorama stitching algorithm that uses cross-dataset nearest neighbours to tackle dataset heterogeneity [21]; scVI uses hierarchical Bayesian model with conditional distributions specified by deep neural networks to aggregate information across similar cells and genes, and to approximate the distributions that underlie observed expression values, which generates a model removing batch effects and retaining the biological signals [22]; SeuratV4 integration uses canonical correlation analysis (CCA) or reciprocal PCA (RPCA) to identify anchors between datasets then uses dynamic time warping to align the subspaces [16].

Despite that a variety of algorithms is available, several challenges specific to cross-species integration make it exceptionally difficult. On the one hand, global transcriptomic difference between species can be much stronger than the average batch effects [17]. It is not uncommon that moderate methods fail to integrate the data [17,23]. On the other hand, species-specific populations may become obscure due to overfitting, especially for nearest-neighbour based algorithms [15]. Last but importantly, cross-species integration requires matching genes from different species by gene homology. Currently, most studies only use one-to-one orthologous gene from public resources such as ENSEMBL[3,12] or EggNOG [9], while some others manually add relevant one-to-many orthologs [16] or use an in-house pipeline [13] for homology mapping. Nevertheless, gene homology relationships can be obscure for species without a well-annotated genome and challenging to map for evolutionally distant species pairs. Recently, a novel method, SAMap, tackles this issue by reciprocally and iteratively update a gene-gene mapping graph and a cell-cell mapping graph to stich atlases between species [23]. SAMap can lay much stronger mapping of homologous cell types across distantly species and is capable of discovering gene paralog substitution events. Nevertheless, SAMap is computationally intensive and designed for whole-body analysis, while methods optimized for large-scale integration may be more scalable to multiple datasets and more species.

Overall, it remains obscure which strategy for cross-species integration can generate candid results. It is also unclear how the selection of homologous genes and integration algorithms affect the observable relationships between cells. Different strategies haven't been tested in a variety of organs, species and biological systems. To date, there is no benchmarking of the data integration strategies for cross-species analysis.

To examine the output characteristics of different single-cell transcriptomics data integration strategies in the cross-species analysis, we assembled BENchmarking inteGration strAtegies for cross-species anaLysis of single-cell transcriptomics data (BENGAL) pipeline. The pipeline is implemented with Nextflow DSL2 [24] and singularity containers [25] for portability and reproducibility, and is compatible with the execution on high performance computing (HPC) clusters. We ran BENGAL to examine 25 cross-species integration strategies, covering 4 ways of mapping genes by homology and 9 integration algorithms in a variety of biological systems. Based upon known cell type homology from published literature, we demonstrate the integration strategy that generates candid output suitable for downstream analysis. Specifically, we investigated: 1) how well can one-to-one homologous cell types be integrated between species quantified by batch metrics; 2) how much cell type distinguishability and biological heterogeneity is lost due to integration quantified by biology conservation metrics; 3) whether it is possible to perform cell type annotation transfer using the integrated data. By observing the behaviour of different integration strategies under certain tasks, we provide guidelines on choosing the most appropriate strategy for different cross-species integration goals.

Results

Integration and assessment pipeline for cross-species scRNA-seq data

The BENGAL pipeline performs cross-species integration and assessment of integration results using 25 strategies, as illustrated in Fig. 1. Quality control and curation of cell ontology annotation of input data is required prior to running the pipeline. During integration, the pipeline first translates orthologous genes between species using ENSEMBL multiple species comparison tool [26] and concatenate raw count matrices. We experimented three ways to handle gene homology mapping cross-species: a mapping using only one-to-one orthologs, and mappings including one-to-many or many-to-many orthologs by selecting those with a high average expression level, or with a strong homology confidence (see Methods). Notably, LIGER UINMF also takes unshared features, therefore genes without annotated homology are added on top of the mapped genes. SAMap requires *de-novo* reciprocal BLAST to construct a gene-gene homology graph, so we followed the method's standalone workflow. We feed the concatenated raw count matrix to 9 top-performing methods from previous benchmarking of data integration [17], including fastMNN, harmony, LIGER, LIGER UINMF, SAMap, scanorama, scVI, SeuratV4CCA and SeuratV4RPCA.

A key premise required to examine the quality of integration output is high-quality annotation of cell types and confident cell type homology between species. In this study, we focus our analysis on the known one-to-one homologous cell types in vertebrates (see Methods for curation of cell type annotation based on known homology). We reason that only if one-to-one homologous cell types are well-integrated, it is then possible to perform deeper clustering analysis to identify cell types at a higher granularity or to decipher complex cell type taxonomy. Hence, we expect well-integrated data to closely embed known one-to-one homologous cell types across species.

We assessed the integration from two main aspects: between-species mixing and biology conservation, by computing and comparing 4 batch effect evaluation metrics or 6 biology conservation metrics, respectively (see Methods). Importantly, we developed a new biology conservation metric, Accuracy Loss of Cell type Self-projection (ALCS), to target the most unwanted artefact of cross-species integration. ALCS quantifies the degree of blending between cell types per-species after integration, thus indicates the tendency of overcorrection of cross-species heterogeneity that may obscures species-specific cell types. We employed the self-projection concept in machine learning implemented in our previous work Single Cell Clustering Assessment Framework (SCCAF) and use the test accuracy of self-projection to indicate how well cell type labels are distinguishable from each other on a given feature representation. The loss of self-projection accuracy after integration, compared with unintegrated per-species data, measures the loss of cell type distinguishability due to integration (see Methods).

Benchmarking metrics of different integration strategies varies widely

We ran the BENGAL pipeline in 7 cross-species integration tasks spanning a variety of scenarios to observe each strategy's relative level of species mixing and biology conservation (Table 1). We used 4 reference tasks to calculate the average species mixing score, biology conservation score, integrated score and ranking for each strategy (Fig. 2). Strategies that achieve successful integrations stayed largely consistent across tasks while the relative species mixing and biology conservation scores varied (Extended Data Fig. 1-4). Generally, scVI, SeuratV4 CCA, SeuratV4 RPCA, harmony and SAMap performed well across the board. scVI showed stronger species mixing while SAMap performed better with respect to biology conservation.

Integration of human and mouse pancreas, and human, macaque and pig hippocampal-entorhinal system

We demonstrate BENGAL, starting from the integration between human and mouse pancreas [27] (Fig. 3A), and human, rhesus macaque and pig hippocampal-entorhinal system [28] (Fig. 3D). These two scenarios have clearly distinguishable cell types and evident one-to-one cell type homology. Meanwhile, the data were generated in the same study thus reducing technical batch effects. In comparison with the pancreas task, the hippocampus task further challenges the strategies with larger data size and more complex subcluster structures (Table 1). Prior to integration, cells clustered strongly by species of origin (Fig.3 B, E). After integration and assessment, we observed diverse outputs of the integration algorithms (Fig. 3C, F, Extended Data Fig. 1, 2). Overall, the integrated scores ranged from 0.18 to 0.90 (pancreas task) or 0.33 to 0.88 (hippocampus task). Batch correction score was between 0.25 to 0.86 or 0.37 to 0.89, and biology conservation score from 0.13 to 0.99 or 0.04 to 0.99, for the pancreas or hippocampus task, respectively. In both cases, harmony, fastMNN, scVI, SeuratV4 CCA, SeuratV4 RPCA and SAMap were able to integrate the data while scanorama, LIGER and LIGER UNIMF did not integrate some homologous cell type pairs. Interestingly, after data integration, only scVI preserved observable substructures within each cell type cluster (*e.g.* CA1 SUB, CA2-3, InN) in the hippocampus task (Fig. 3F). SAMap gave a strong cross-species alignment, overlapping almost all shared cell types.

Integrating embryonic development between xenopus and zebrafish

Besides integrating data within one organ among adult animals, it is also of great interest to integrate the whole-body embryonic development process. In this scenario, cell types are less distinct and a continuous trajectory exists among cells from different developmental stages. Based on mapped homologous cell types and recorded hours post fertilization (HPF) between a zebrafish and xenopus embryonic development dataset [29,30], we examined cross-species mixing and biology conservation of

various strategies' integration output. We observed that SAMap significantly outperformed others in this task with exceptional biology conservation largely attribute to a high ALCS (integrated score of SAMap: 0.85, others: 0.83-0.33, Extended Data Fig. 4). Interestingly, as seen on the SAMap embedding, homologous cell types were strongly linked to each other and cells from different developmental stages were partially mixed. Whereas, for other methods such as scVI and SeuratV4, the progression of developmental stages along the embedding was more explicit (Extended Data Fig.10). This is in line with a relatively low trajectory conservation score in SAMap (Extended Data Fig.4). This observation suggests that SAMap can prioritize identity-defining expression characteristics that stitch together evolutionarily related cell types.

Cross-atlas integration of heart and aorta among five species

We designed a challenging task that is to integrate the heart and aorta tissue from five species: human, long-tail macaque, mouse, xenopus and zebrafish (Fig. 4A). While the mouse data was from an independent study to cover more cell types [31], all other species data were from current multi-organ atlases [1,3,5,32]. We performed the integration by gradually adding one more species, starting from human and long-tail macaque, to explore the upper limit of the evolutionary distance across species at which integration is still possible (Fig. 4B).

When comparing cell type annotations between atlases, we noticed that the annotation granularity varies greatly (Extended Data Fig. 5-9). To robustly match homologous population between atlases based on existing expert annotation, we sought help from cell ontology (CL), a controlled vocabulary system that describes the ancestor-descendant relationship of cell type terms. CL is the current standard vocabulary used in both the Human Cell Atlas [33] and the EBI Single Cell Expression Atlas [34] for cell type annotation. Leveraging the hierarchical organization of CL, we developed scOntoMatch, an R package which aligns the granularity of ontology annotations of scRNA-seq datasets to make them comparable across studies (see Methods for details). Applying scOntoMatch to the heart data, we found the layer of cell type annotation granularity that was shared across any given group of species.

Prior to integration, cell types were highly distinguishable from each other for all species, indicated by a test accuracy of self-projection over 0.94 (Fig. 5A). After integration, scVI gave the highest species mixing (batch score 0.83-0.7 for four scenarios, unscaled within each scenario, Fig. 4C), followed by LIGER (0.73-0.63), SeuratV4 CCA (0.71-0.66), SeuratV4 RPCA (0.66-0.64), harmony (0.66-0.42), fastMNN (0.64-0.57), LIGER UINMF (0.62-0.57) and scanorama (0.54-0.49). Methods achieved good biology conservation including SeuratV4 CCA (biology conservation score 0.99-0.96 for four scenarios), harmony (0.98-0.95), SeuratV4 RPCA (0.97-0.95), scVI (0.96-0.82), scanorama (0.95-0.90) and fastMNN (0.94-0.91), while LIGER UINMF (0.80-0.08) and LIGER (0.66-0.03) experienced strong loss of biology. With the addition of species, the relative performance of different methods stayed

stable and batch scores per method did not decrease significantly, suggesting that methods could stay robust when involving more datasets.

In contrast, biology conservation scores dropped and cell types became more blurred when distant species were added to the integration (Fig. 4C). Taking scVI integrated one-to-one orthologs only data as an example. While integrating between human and macaque or human, macaque and mouse, cell type distinguishability per species stayed the same after integration. Adding xenopus to the integration, xenopus data experienced some loss of cell type distinction, while the loss became global and stronger after zebrafish was integrated with the rest species. The loss of cell type distinguishability was also broadly observed in other strategies shown in Fig. 4E. The results suggest that due to the information loss during gene homology mapping and the integration process, integrating distant species cannot generate candid results suitable for downstream analysis such as de-novo clustering.

Cell type annotation transfer cross-species

One key application of cross-species integration is to perform annotation transfer: using annotated cells in a well-studied species to label the cell types of a newly characterized species. To investigate different strategy's potential to serve as basis of cross-annotation, we performed annotation transfer of shared cell types for all pairs of species using the integrated embedding. In practise, we apply the SCCAF multinomial logistic classifier trained on one species to predict the cell type labels of another species. When the transferred labels cover all known sharded cell types and have an acceptable adjusted rand score (ARI) with the original label, we consider cross-annotation as successful.

Overall, we observed that if data integration was well-achieved, annotation transfer was mostly successful. In the pancreas and heart task, SAMap, scVI, SeuratV4 methods and harmony achieved successful transfer, while for the hippocampus task it was SAMap, scVI and SeuratV4 methods. Nonetheless, minor errors can occur between highly similar cell types. For instance, in the SeuratV4 CCA integrated heart data between human, long-tail macaque and mouse, slight confusion appears between pericyte and smooth muscle cells for long-tail macaque model annotated human data, as well as between fibroblast and smooth muscle cells for mouse model annotated human data (Fig. 5A). Interestingly, we found that SeuratV4 methods integrated data could serve as a basis for a successful label transfer even between evolutionarily distant species such as zebrafish and human (Fig. 5B), despite that some errors were found in the fibroblast population. We noticed that with the addition of evolutionarily more distant species, the annotation transfer quality generally decreases (Fig. 5C).

Impact of one-to-many and many-to-many homologous genes on integration performance

We observed that only in the xenopus zebrafish development task, methods achieved successful integration benefited from including one-to-many and many-to-many orthologs, especially when the

in-paralog with higher homology confidence is matched with their respective ortholog (0.07 ± 0.04 increase of biology conservation score, paired sample Wilcoxon signed-rank test P-value=0.03, Extended Data Fig. 4, Extended Data Fig. 11). We attribute this phenomenon to the high number of one-to-many and many-to-many orthologs gene count between xenopus and zebrafish (Table 1). Adding these in-paralogs gives a more complete picture of cell type expression profile which contributes to better data integration. This observation is also in line with the result that SAMap significantly outperforming others in the embryo task. For the heart task, adding in-paralogs did not help when xenopus and zebrafish are integrated with mammals, which we reason was due to lacking confident one-to-many and many-to-many orthologs among the five species (Table 1).

Discussion

We benchmarked 25 strategies for cross-species single-cell transcriptomics data integration, covering 4 ways to match genes cross-species by homology and 9 integration algorithms. Based on known homologous cell types, we used 4 metrics to evaluate the degree of species mixing and 6 for scoring biology conservation. We also developed a novel metrics, ALCS, to measure how well cell type distinguishability is preserved after cross-species integration. Generally speaking, we found that scVI achieves an overall balance between species mixing and biology conservation, while SeuratV4 methods can perform strong integration even among evolutionarily distant species. We noticed that the tailor-made cross-species integration algorithm, SAMap, can highlight core similarities between homologous cell types that are otherwise less discernible. By performing annotation transfer on shared cell types cross-species using integrated results, we observed that if complete integration is achieved, annotation transfer can be to a large extent successful even between evolutionarily distant species such as human and zebrafish. Nonetheless, cell types highly similar with each other can still be mislabelled between closely related species *e.g.* human and mouse.

Based on a diversity of cross-species integration scenarios investigated in this study, we provide the following guidelines on choosing the most appropriate algorithm for cross-species integration: for closely-related species, scVI or harmony carries out species mixing while preserving biological heterogeneity. For evolutionarily distant species, SeuratV4 methods can achieve strong species mixing and RPCA is more scalable than CCA for larger datasets. For integration of whole-body atlases or between species lacking well-curated gene homology annotation, SAMap excel in aligning homologous cell types by resolving the gene homology mapping challenge. For species sharing a large amount of one-to-many and many-to-many orthologs, including them in the analysis can improve the integration as more information about cell type expression profiles are preserved.

Our quantitative scoring system of species mixing and biology conservation are appropriate evaluations under two assumptions: 1) cell type annotations are accurate; 2) known cell type

homology from literature is confident. It is important to point out that our knowledge about cell types and their evolution is still rapidly growing [8,35]. Currently, cell type homology is discernible at decreased granularity with the increase of evolutionary distance between species due to both biological and practical reasons. Biologically speaking, more complex species can have a greater number of specialized cell types while all of them collectively are homologous to one multi-function cell type in their evolutionary ancestors. In practice, cell type annotation can only reach a relatively coarse level in less-studied species and tissues due to our limited understanding. In this study, we focus on one-to-one homologous cell types and expect them to mix between species and stay distinctive among other cell types after integration. We demonstrated that leveraging the unified CL annotation system, we can find the annotation granularity at which one-to-one cell type mapping is established in homologous organs among vertebrates.

Here we further address several future paths for cross-species integration. To date, cross-species integration relies heavily on sequence orthology. Such an approach inherently assumes orthologs to have the same expression pattern and functions to distinct cell types, which has two main caveats: 1) species-specific genes contribute to a large extent the novelty of species-specific cell types and shall not be ignored when performing cross-species analysis [11,36], 2) functional diversification between orthologs might be as common as between paralogs and out-paralogs can exhibit higher expression similarity compared with their corresponding orthologs through gain, loss and partition of function [23,37,38]. Moreover, expression similarity of cells cross-species may arise from several reasons including homology, convergence and concerted evolution [8]. It is important to disentangle the source of the shared expression features to achieve deeper understanding of cell types [10]. Finally, one-to-one homologous relationship between cell types might not hold for evolutionarily distant species, as cell types have undergone complex evolution of identity-defining gene networks (see the interconnected cell type groups across Metazoa demonstrated in the SAMap study [23]). According to the heart example, we conclude that when data integration is performed between non-mammals with mammals, the result can still serve as basis for cell type annotation transfer, but the embedding might be not suitable for de-novo clustering analysis due to strong loss of biology. Other analysis, such as expression correlation based independent analysis might be more informative [9,13]. On balance, novel integration strategies aware of evolutionary homology and adopt a hierarchical structure of cell type classification will provide deeper insights for cross-species analysis.

We anticipate that the computational tools developed in this study will aid future research to obtain the most insightful cross-species integrations by selecting the optimal strategy. We provide BENGAL, a Nextflow pipeline for cross-species scRNA-seq data integration and assessment of integration results. We propose ALCS as a cross-species integration-focused biology conservation metric to quantify the loss of cell type distinguishability. To align the granularity of cell ontology annotation

across datasets, we developed an R package scOntoMatch. One key use case is to integrate between several species densely sampled in a clade to reconstruct cell type evolution. Another valuable application is to compare cell types between human and model animals, with an aim of providing molecular guidance to select the most appropriate model for a particular tissue or disease. While more and more parallel biological processes, such as embryonic development [29,30,39,40] and aging [41] being characterized in several species at the cellular level, it is an exciting time to extend our knowledge base of cells, cell types and tissue functions in the area of development, evolution, health and disease. To obtain multi-modal data, cellular lineage information and cellular response under perturbation then compare cross-species will give a more comprehensive picture of cellular profile to deeper our understanding of the history, identity, and the behaviour of cells.

Online Methods

Datasets used in this study

We downloaded raw count matrices and published annotations of the following datasets: inDrop data from human and mouse pancreas (GSE84133) [27], snRNA-seq data from human, macaque and pig hippocampus (GSE186538) [28], heart and aorta tissue from human (https://figshare.com/projects/Tabula_Sapiens/100973) [1], long-tail macaque (https://db.cngb.org/nhpca/download) [3] and mouse (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-8810) (only no compound treatment mouse data is used) [42] and inDrops data from zebrafish (GSE112294) [29] and xenopus embryo (GSE113074) [30]. We use the published cell type annotation of all these datasets and exclude cells that didn't pass quality control in the original study. For the hippocampus task, Neuroblasts, radial glia-like cells and neural intermediate progenitor cell from macaque and pig are collectively annotated as "neuronal progenitor". In the human data, we collectively annotate CA1 and SUB as "CA1 SUB", CA2, CA3 as "CA 2-3", macrophage, microglia, myeloid cell and T cell as "immune", arterial and ventricle smooth muscle cell as "smooth muscle cell", pericytes and vascular and leptomeningeal cell as "vasculature". For the zebrafish and xenopus embryo task, we used the annotations provided in the SAMap study [23], as they have manually matched one-to-one homologous cell types using multiple lines of evidence.

Transcriptomes used in the BLAST step in SAMap are downloaded from ENSEMBL, except that for the Xenopus in embryo task is downloaded from Xenbase (https://www.xenbase.org/entry/static-xenbase/ftpDatafiles.jsp). Versions are in line with the version at original publication of each dataset: pancreas task: mouse (GRCm38) and human (GRCh38); hippocampus task: human (GRCh38), macaque (Mmul_10), pig (Sscrofa11); heart task: human (GRCh38), monkey

(Macaca_fascicularis_6.0) and mouse (GRCm39); embryo task: Xenopus (Xtropicalisv9.0), zebrafish (GRCz10).

Curation of cell type annotation based on known homology

We paid specific attention that the same cell type identifier refers to homologous cell types *i.e.* cell types from different species and trace back to the same precursor cell type in a common evolutionary ancestor [8]. In the pancreas and hippocampus tasks, cell type homology is evident between species and matched in the original study. In the embryo task, we used a curated annotation from the SAMap study, which used a combination of the manual matching and developmental hierarchies. For the heart task, we leverage the hierarchical structure of cell ontology to robustly align annotation granularity that matches homologous cell populations for a given set of species with curated ontology annotation (see the 'Aligning ontology annotation with scOntoMatch' section below). Overall, we ensure one-to-one mapping of homologous cell types among the studies species.

The BENGAL pipeline

The BENGAL pipeline is built using Nextflow (v22.04.3) DSL2 in java (OpenJDK v11.0.9.1-internal). For data integration, python-based scripts were built based on Python (v3.9.13), Scanpy (v1.9.1), h5py (v3.7.0) and anndata (v0.7.5). We used python implementation of harmonypy (v0.0.5), scanorama (v1.7.2), scVI (v0.15.0 with pytorch v1.12.1 and cudatoolkit v11.6 to support execution with Nvidia GPU), SAMap (v1.0.2). R-based scripts are based on R (v4.0.5). We used Seurat (V4.1.1), LIGER (v0.5.0) and LIGER UINMF (v1.1.0) and fastMNN (v1.12.3 from package batchelor). For SCCAF analysis, we modified the original SCCAF package and deposed our version at https://github.com/YY-SONG0718/sccaf. For batch metrics and biology conservation metrics calculation, we used scIB (v1.0.2).

Translating features cross-species by gene homology

We used ENSEMBL multiple species comparison tool (version 106), accessed via biomaRt (v2.46.3), to annotate gene homology and translate ENSEMBL gene id cross-species. We chose ENSEMBL to be the primary database for gene homology annotation, since there is manually curated homology, in addition to the automated annotation pipeline, it is a popular choice among previous and current cross-species scRNA-seq studies and it is broadly accessible for biologists from various disciplines. BENGAL is conveniently transferrable to use other homology databases, such as EggNOG or BLAST, as long as a homology matching table can be provided.

Aligning ontology annotation with scOntoMatch

We developed scOntoMatch, an R package which aligns the granularity of ontology annotation of scRNA-seq datasets to make them comparable across studies. Cell type terms can be robustly described by cell ontology (CL), a structured controlled vocabulary for cell types in animals [33]. Cell ontology (CL) is a species-neutral ontology systems that describes the relationships between cell type terms, such as ancestor-descendant, and it covers cell types from prokaryotes to mammals.

scOntoMatch performs annotation alignment by matching descendant terms to ancestor terms so that every cell type term is a leaf node in the directed acyclic graph (DAG) created by all cell types and their ancestors from all the input datasets. Taking any number of single cell datasets with ontology labels, the algorithm first trims the cell type tree in each dataset so that there is no ancestor-descendant relationship between cell types within the dataset. The main step is to identify an ontology matching among the datasets, in which direct matching and mapping from descendants to ancestors are performed to find the last common ancestor (LCA) term. The final output is an ontology annotation that is aligned across datasets.

Construct concatenated count matrix cross-species

Starting from raw count matrices of scRNA-seq data from different species, the BENGAL pipeline first identifies gene groups composed of one-to-one, one-to-many or many-to-many orthologs and species-specific genes using ENSEMBL multiple species comparison tool. One-to-one orthologs are directly matched across species. On top of that, one-to-many or many-to-many orthologs within each gene group are added and matched by selecting those with higher average expression level calculated using scanpy.pp.calculate_qc_metrics, or with stronger homology confidence defined in ENSEMBL using available attributes among: orthology.confidence, Gene.order.conservation.score, Whole.genome.alignment.coverage, query.gene.identical.to.target, gene.identical.to.query.gene in biomaRt. The method LIGER UINMF also takes unshared features (species-specific genes) and these genes were added in addition to the three types of inputs with shared features. SAMap quantifies gene homology strength using de-novo BLAST, so the raw count matrix and the full set of genes are given to the method without the above homology-matching step, and tblastx (from BLAST v2.12.0) was run using transcriptomes data to generate blast maps for the method.

Standard processing of scRNA-seq data

Per-species data and unintegrated data are analysed following the scanpy standard analysis workflow with following parameters: sc.pp.highly_variable_genes(adata, min_mean=0.0125, max_mean=3, min_disp=0.5); sc.pp.scale(adata, max_value=10); sc.tl.pca(adata, svd_solver='arpack'); sc.pp.neighbors(adata, n_neighbors=15, n_pcs=40); sc.tl.umap(adata, min_dist=0.3, spread=1).

Dimension reduction and visualization

For methods output an embedding, neighbourhood graphs are calculated with n_neighbors=15, n_pcs=40, and the Uniform Manifold Approximation and Projection (UMAP) representation is computed using min_dist=0.3 and spread=1.0. For methods output pseudo-count matrix, we first calculate PCA, then use the same parameters to calculate neighbourhood graph and UMAP. SAMap output a corrected neighbourhood graph, so we only recalculate UMAP using the same parameter for harmonizing visualization.

Cross-species data integration

Taking the concatenated matrix with cells from all species and features matched by homology as input, the pipeline runs 9 integration methods using default parameters.

fastMNN: we use the fastMNN from github.com/LTLA/batchelor, via SeuratWrappers (v0.3.0) following the tutorial http://htmlpreview.github.io/?https://github.com/satijalab/seurat-wrappers/blob/master/docs/fast_mnn.html.

harmony: we use scanpy.external.pp.harmony_integrate to run harmony following tutorial: https://scanpy.readthedocs.io/en/stable/generated/scanpy.external.pp.harmony_integrate.html

LIGER: we use the LIGER method via SeuratWrappers (v0.3.0) following the tutorial: http://htmlpreview.github.io/?https://github.com/satijalab/seurat-wrappers/blob/master/docs/liger.html

LIGER UINMF: we run LIGER UINMF via github.com/welch-lab/liger following the tutorial http://htmlpreview.github.io/?https://github.com/welch-lab/liger/blob/master/vignettes/SNAREseq_walkthrough.html

SAMap: we run SAMap using github.com/atarashansky/SAMap following the tutorial and vignette in the github repository.

Scanorama: we use scanpy.external.pp.harmony_integrate to run scanorama following tutorial: https://scanpy.readthedocs.io/en/stable/generated/scanpy.external.pp.scanorama_integrate.html

scVI: we run scVI from scvi-tools.org/ following tutorial https://docs.scvi-tools.org/en/stable/tutorials/notebooks/harmonization.html

SeuratV4CCA and SeuratV4RPCA: we run SeuratV4CCA from https://satijalab.org/seurat/index.html following https://satijalab.org/seurat/articles/integration_introduction.html and RPCA following https://satijalab.org/seurat/articles/integration_rpca.html

SAMap is designed to perform integration between data from different species so 'species' is used as integration key. For the integration methods other than SAMap, we use 'species' as integration key if

per-species data do not appear to have strong batch effect or per-species batches are unbalanced in terms of cell type composition (heart task), otherwise, we use batch as integration key if each batch have balanced cell types (pancreas and hippocampus task). SeuratV4 methods output a pseudo-count matrix, SAMap output a cell-cell neighbourhood graph, while other methods output an embedding. We then feed the output from different methods to the assessment framework.

Batch removal metrics

Species mixing is assessed by the relative reduction of batch metrics before and after integration. We compute the principal component regression (PCR) score, batch average silhouette width (ASW) score, graph connectivity (GC) score and batch local inverse Simpson's index (LISI) score using the scIB framework. All metrics are computed on the PCA embedding for methods returning a pseudo-count matrix and the output embedding for methods returning an embedding. We did not use kBET, since cross-species batch effect is strong while kBET being highly sensitive. In all the tasks, kBET score per cell type stays highly significant after integration, thus cannot inform the relative performance of different strategies.

(1) PCR score: assume that there is batch effect in the dataset, then the heterogeneity caused by batch shall be captured by some principal components (PC) in the PC representation of the dataset. Regression of the batch covariant will then significantly correlate with some PCs. The sum of all variance contribution by batch from all PCs (Var(C|B) gives an approximation of the level of batch effect in the dataset. PCR score is the scaled Var(C|B) before and after data integration.

$$Var(C|B) = \sum_{i=1}^{G} Var(C|PC_i) \times R^2\,(PC_i|B)$$

$$PCR = \frac{Var(C|B)_{pre} - Var(C|B)_{post}}{Var(C|B)_{pre}}$$

PCR range from 0 to 1, with 0 means no change of batch contribution and 1 means complete removal of batch contribution after integration.

(2) Batch ASW: silhouette measures the relationship between with-in cluster distance of a cell and the between-cluster distance of that cell to the closest cluster. The batch ASW is computed based on the absolute silhouette of batch labels per cell i of cell type j, then averaged across M cell types.

$$bASW = \frac{1}{|M|}\sum_{j\in M}\frac{1}{|C_j|}\sum_{i\in j}(1 - |s(i)|)$$

Batch ASW range from 0 to 1, with 0 meaning strongly separated batches and 1 meaning ideal mixing of batch.

(3) GC: the graph connectivity score measures whether the k-nearest neighbour (KNN) representation of the data connects all cells with the same label. GC is measured by the fraction of cells belongs to the largest connected component (LCC) of the subgraph constructed with all cells i with the same label j, averaged across M cell types.

$$GC = \frac{1}{|M|}\sum_{j\in M}\frac{|LCC\big(G(N_c; E_c)\big)|}{|C_j|}$$

GC range from 0 (exclusive) to 1, with the lowest possible score meaning entire separation of cells with the same label in the graph and 1 meaning complete connection between cells with same label.

(4) Batch LISI score: LISI is a measurement of the degree of concentration when individuals are classified into types. The LISI measure equals the number of random draws needed to first get two cells from a neighbour list that are from the same group. If we use batch as group, well mixed data will have a higher LISI.

$$LISI = \frac{1}{\sum_{i=1}^{R} p_i^2}$$

As LISI range from 1 to number of batches, we scale the LISI to 0 to 1, with 0 meaning low batch integration and 1 meaning highly mixed batch. Since LISI score depend on the neighbourhood size, we use the graph LISI implemented in scIB to ensure methods with different output types are comparable with each other [17]. For methods output pseudo-count matrix, we calculate KNN using first 20 PCs on PCA embedding, for SAMap, we use its original KNN output, for methods output an embedding, we calculate KNN using first 20 dimensions on the output embedding.

Assessing biology conservation with ALCS

We adapted the Single Cell Clustering Assessment Framework (SCCAF) [43] to calculate ALCS, a biology conservation metric indicative of loss of cell type distinguishability due to data integration. SCCAF is a self-projection-based approach to assess the validity of a classification system. For cells with given labels on a feature embedding, we first split the cells into a training set and a test set for each class. A machine learning classifier is then trained on the training set and used to predict the

label of the test set. In our case, we chose to use multinomial logistic classifiers as it has been shown to perform equally well compared with other models in the SCCAF study. The accuracy of self-projection is used to quantify the clarity of the classification. We compute the loss of self-projection accuracy on the integrated embedding compared with the original per-species data, as a measurement of how much cellular specificity is lost due to integration. SCCAF models are trained on the PCA embedding for methods returning a pseudo-count matrix and the method output embedding for the rest methods.

Biology conservation metrics

We calculate cLISI, cell type silhouette, NMI, ARI and trajectory conservation as biology conservation metrics using the scIB framework.

(1) Cell type LISI (cLISI): If cell type label is used as groups in LISI score calculation, well separated cell types in data will lead to a lower LISI. We reversely scale cLISI to a value between 0 and 1 where 0 indicates cell type mixing and 1 means cell types are kept distinguishable.
(2) Cell type average silhouette width (cASW): the cASW is to scale the ASW of cell types to a score between 0 and 1. 0 means there is higher intra-cluster heterogeneity and 1 means well separated, clear cell types.

$$cASW = \frac{(ASW_c + 1)}{2}$$

(3) Normalized mutual information (NMI): describes the similarity of two clustering system. Here we compare cell type annotation with optimized Louvain clustering as described in the scIB framework [17]. Louvain clustering was performed at a resolution range of 0.1 to 2 in steps of 0.1, and the clustering output with the highest NMI with the label set was used. NMI is scaled between 0 or 1 correspond to uncorrelated clustering or a perfect match.
(4) Adjusted rand index (ARI): quantifies the overlap of two labelling systems. We calculate ARI between annotation and optimized Louvain clustering, same with the clustering used in (3) NMI via the scIB framework [17]. ARI is scaled between 0 or 1 correspond to random labelling or a perfect match.
(5) Trajectory conservation score: we used the diffusion pseudotime implemented in scanpy.tl.dpt with default parameters to calculate pseudotime coordinate of cells. The dpt score of each cell is given as input to scIB to compare between pre and post integration for each species.

Data and codes availability

The BENGLE pipeline is available at https://github.com/Functional-Genomics/CrossSpeciesIntegration. Codes and source data for generating the figures in this study are deposed at https://github.com/Functional-Genomics/BENGAL_reproducibility. The package scOntoMatch is available through CRAN https://cran.r-project.org/web/packages/scOntoMatch/index.html, which works on Seurat objects, or https://github.com/Functional-Genomics/scOntoMatch which have two versions that operates on AnnData objects or Seurat objects.

References

1. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

2. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

3. Han, L. *et al.* Cell transcriptomic atlas of the non-human primate Macaca fascicularis. *Nature* **604**, 723–731 (2022).

4. Wang, F. *et al.* Endothelial cell heterogeneity and microglia regulons revealed by a pig cell landscape at single-cell level. *Nat. Commun.* **13**, 1–18 (2022).

5. Jiang, M. *et al.* Characterization of the Zebrafish Cell Landscape at Single-Cell Resolution. *Front Cell Dev Biol* **9**, 743421 (2021).

6. Musser, J. M. *et al.* Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science* **374**, 717–723 (2021).

7. Vergara, H. M. *et al.* Whole-body integration of gene expression and single-cell morphology. *Cell* **184**, 4819-4837.e22 (2021).

8. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).

9. Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).

10. Woych, J. *et al.* Cell-type profiling in salamanders identifies innovations in vertebrate forebrain evolution. *Science* **377**, eabp9186 (2022).

11. Shafer, M. E. R., Sawh, A. N. & Schier, A. F. Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts. *Nat Ecol Evol* (2021) doi:10.1038/s41559-021-01580-3.

12. Li, H. *et al.* Cross-species single-cell transcriptomic analysis reveals divergence of cell composition and functions in mammalian ileum epithelium. *Cell Regen* **11**, 19 (2022).

13. Xu, J. *et al.* Transcriptional and functional motifs defining renal function revealed by single-nucleus RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2203179119 (2022).

14. Liang, C., Musser, J. M., Cloutier, A., Prum, R. O. & Wagner, G. P. Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. *Genome Biol. Evol.* **10**, 538–552 (2018).

15. Shafer, M. E. R. Cross-Species Analysis of Single-Cell Transcriptomic Data. *Front Cell Dev Biol* **7**, 175 (2019).

16. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

17. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* (2021) doi:10.1038/s41592-021-01336-8.

18. Zhang, F., Wu, Y. & Tian, W. A novel approach to remove the batch effect of single-cell data. *Cell Discov* **5**, 46 (2019).

19. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

20. Kriebel, A. R. & Welch, J. D. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.* **13**, 1–17 (2022).

21. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

22. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

23. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10**, (2021).

24. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

25. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).

26. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

27. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360.e4 (2016).

28. Franjic, D. *et al.* Transcriptomic taxonomy and neurogenic trajectories of adult human, macaque, and pig hippocampal and entorhinal cells. *Neuron* (2021) doi:10.1016/j.neuron.2021.10.036.

29. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

30. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).

31. McLellan, M. A. *et al.* High-Resolution Transcriptomic Profiling of the Heart During Chronic Stress Reveals Cellular Drivers of Cardiac Fibrosis and Hypertrophy. *Circulation* **142**, 1448–1463 (2020).

32. Liao, Y. *et al.* Cell landscape of larval and adult Xenopus laevis at single-cell resolution. *Nat. Commun.* **13**, 4306 (2022).

33. Osumi-Sutherland, D. *et al.* Cell type ontologies of the Human Cell Atlas. *Nat. Cell Biol.* **23**, 1129–1135 (2021).

34. Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.* **50**, D129–D140 (2022).

35. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).

36. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323.e30 (2016).

37. Stamboulian, M., Guerrero, R. F., Hahn, M. W. & Radivojac, P. The ortholog conjecture revisited: The value of orthologs and paralogs in function prediction. Bioinformatics 36: i219--i226. (2020).

38. Studer, R. A. & Robinson-Rechavi, M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* **25**, 210–216 (2009).

39. He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760–767 (2020).

40. Calderon, D. *et al.* The continuum of *Drosophila* embryonic development at single-cell resolution. *Science* **377**, eabn5800 (2022).

41. Wang, R. *et al.* Construction of a cross-species cell landscape at single-cell level. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac633.

42. Paik, D. T. *et al.* Single-Cell RNA Sequencing Unveils Unique Transcriptomic Signatures of Organ-Specific Endothelial Cells. *Circulation* **142**, 1848–1862 (2020).

43. Miao, Z. *et al.* Putative cell type discovery from single-cell gene expression data. *Nat. Methods* **17**, 621–628 (2020).

44. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

Acknowledgements

This work was supported by: the European Molecular Biology Laboratory (Y.S., Z.M., A.B., I.P.); the EMBL international PhD program (Y.S.); and the Biotechnology and Biological Sciences Research Council (BBSRC) grant 'Fly Cell Atlas' [BB/T014563/1] (I.P.).

Author contributions

All authors conceived the project. Y.S. designed the framework and curated data, wrote the pipeline and software, performed data analysis and visualization, with supervision from I.P. and inputs from Z.M. and A.B.. Y.S. wrote the initial draft of the manuscript and all authors participated in review and editing.

Ethics declaration

Competing interests: the authors declare no competing interests.

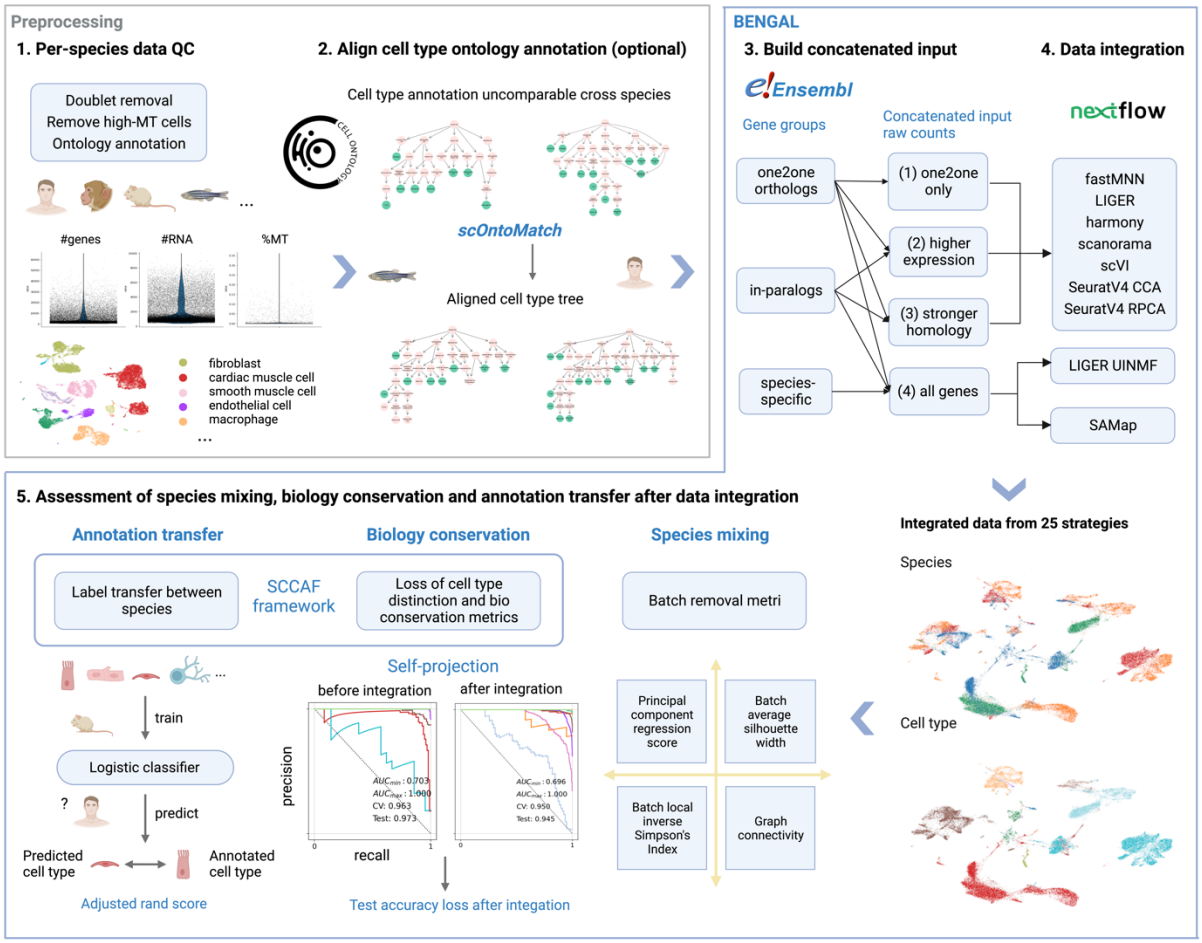


Figure 1 Schematic of BENGAL pipeline. 1) quality control of input data is performed prior to data integration. Potential doublets and low-quality cells expressing high mitochondrial genes are removed. Cell ontology annotation is collected from atlases or data portals, or curated from original published annotation 2) when ontology annotation granularity is incomparable cross datasets, one-to-one homology between cell types need to be robustly aligned. We developed scOntoMatch to find the appropriate annotation granularity and align cell type hierarchies across given datasets (see Methods). 3) genes are grouped and translated across species by homology defined in ENSEMBL multiple species comparison tool. Raw count matrices are then concatenated across species using four possible homology matching methods respective to method inputs. 4) run 9 integration algorithms to generate integrated output. 5) preform integration assessment from species mixing, biology conservation and annotation transfer. BENGAL: BENchmarking inteGration strAtegies for cross-species anaLysis of single cell transcriptomics data; QC: quality control; MT: mitochondrial; SCCAF: single cell clustering assessment framework; AUC: area under curve, CV: cross-validation. Created with BioRender.com.

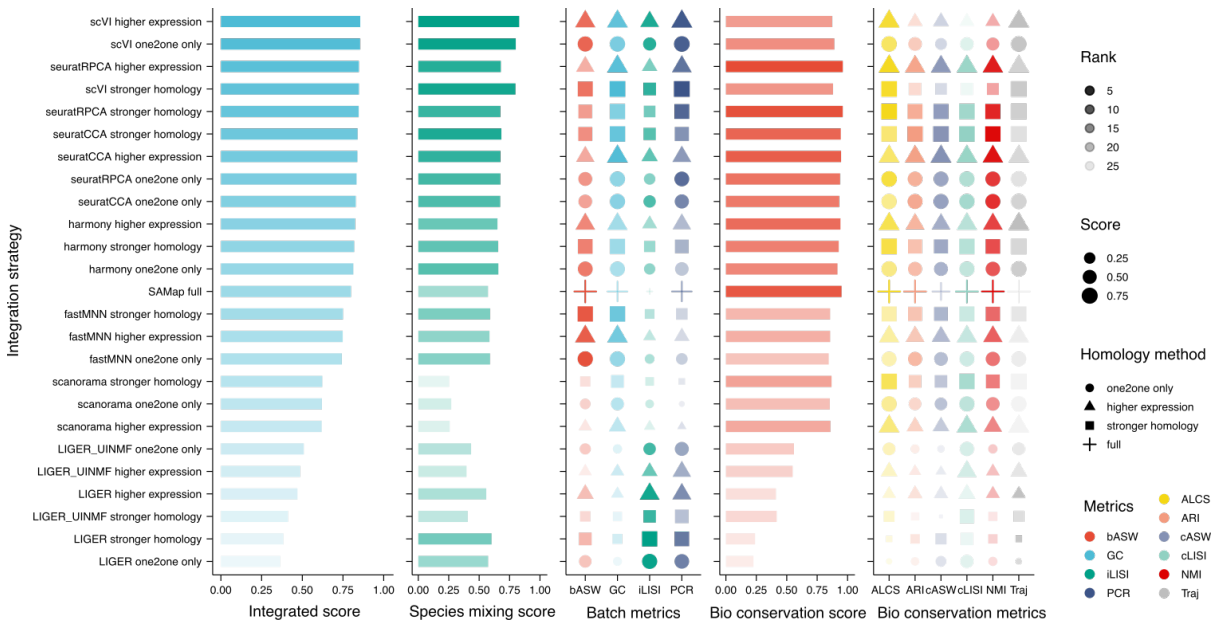


Figure 2 Benchmarking scores and metrics of different integration strategies on cross-species analysis. We use 4 batch metrics and 6 biology conservation metrics to examine the output of 25 data integration strategies on 4 reference cross-species data integration tasks. Integrated score is 0.4*species mixing score and 0.6*biology conservation score. Species mixing score is the average of 4 batch metrics, while biological conservation score is 0.75* ALCS plus 0.05* for the other 5 metrics. Batch metrics and biological conservation metrics are min-max scaled per task and the average scores across 4 tasks are shown in the figure. Trajectory conservation metric is only calculated for the relevant task: integrating embryonic development between xenopus and zebrafish. Bio: biology; bASW: batch average silhouette width; GC: graph connectivity; iLISI: batch local inverse Simpson's index; PCR: principal component regression; ALCS: accuracy loss of cell type self-projection; ARI: adjusted rand index; cASW: cell type average silhouette width; cLISI: cell type local inverse Simpson's index; NMI: normalized mutual information of cell label; Traj: trajectory conservation.

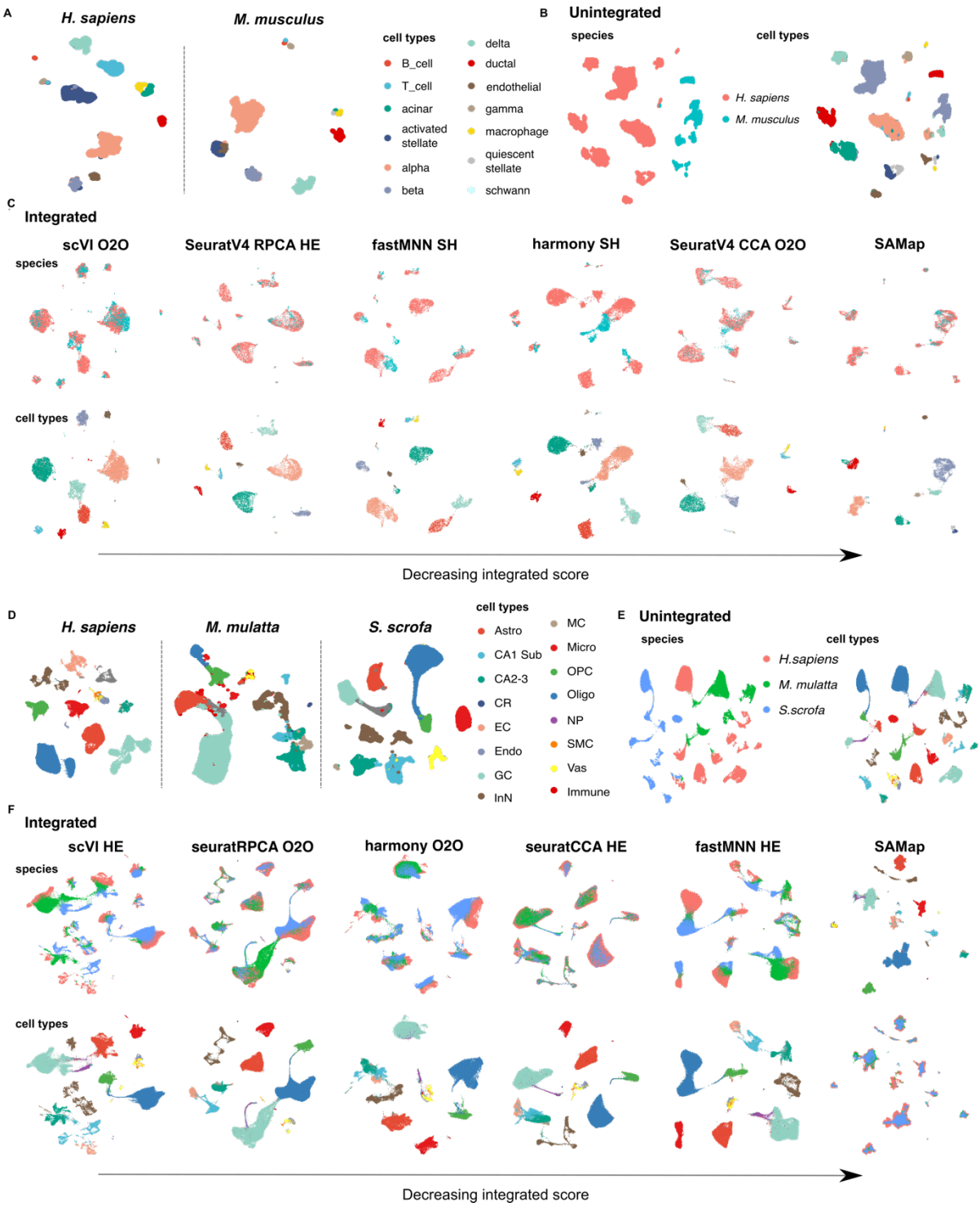


Figure 3 Integration results of human and mouse pancreas and human, macaque and pig hippocampal-entorhinal system. A, UMAP visualization of cell types in human and mouse pancreatic task. B, UMAP visualization of unintegrated data of the pancreas task, in which cells cluster by species and homologous cell types form distinct clusters. One-to-one orthologs between human and mouse were used to concatenate raw count matrices, then processed by standard analysis pipeline (see Methods). C, UMAP visualization of integration results of five top-performing strategies and SAMap of the pancreas task, organized by decreasing integrated score and coloured by species or

cell type. D, UMAP visualization of cell types in human, macaque and pig hippocampal-entorhinal system. E, UMAP visualization of unintegrated data of the hippocampus task generated with the same method used in B. F, UMAP visualization of integration results of five top-performing strategies and SAMap of the hippocampus task, organized by decreasing integrated score and coloured by species or cell type. H. sapiens: Homo sapiens; M. musculus: Mus musculus; M. mulatta: Macaca mulatta; S. scrofa: Sus scrofa; O2O: only use one-to-one orthologs; HE: one-to-one orthologs plus one-to-many and many-to-many orthologs matched by higher average expression level; SH: one-to-one orthologs plus one-to-many and many-to-many orthologs matched by stronger homology confidence; Astro: astrocytes; CR, Cajal-Retzius cells; EC, entorhinal cortex; Endo, endothelial cells; GC, granule cell; InN, inhibitory neurons; MC, mossy cell; Micro, microglia cells, OPC, oligodendrocyte progenitor cells; Oligo, oligodendrocytes; NP, neuronal progenitors; SMC, smooth muscle cells; Vas, vasculature.

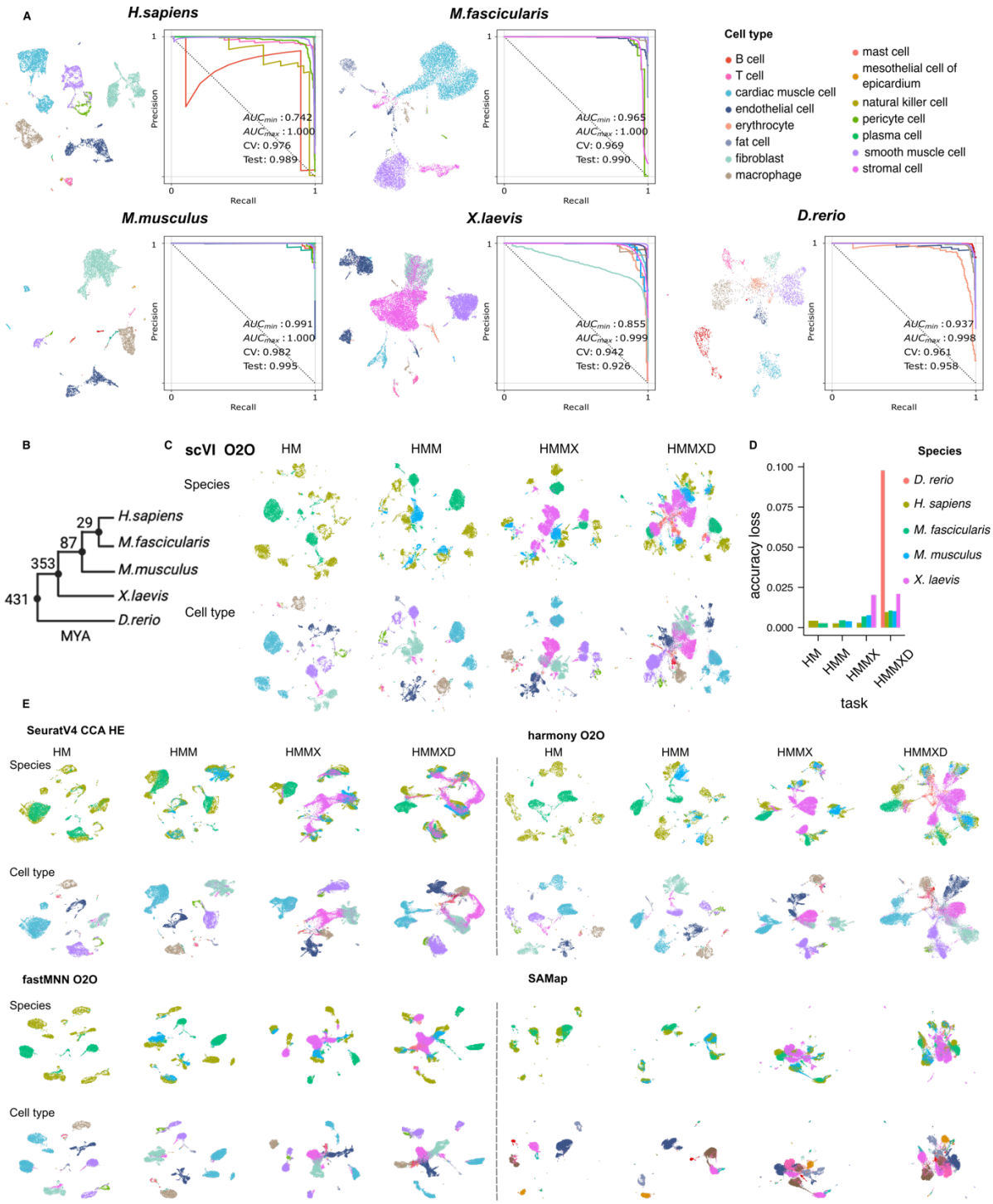


Figure 4 Integrating heart tissue from human, long-tail macaque, mouse, xenopus and zebrafish. A, UMAP visualization of heart data per species coloured by cell types, and the corresponding receiver operating curve of SCCAF assessment results. Prior to integration, cell types are highly distinguishable in each species. SCCAF is a self-projection-based method to examine the validity of cell type classification based on a feature embedding. The test accuracy of SCCAF classifier indicate the level of cell type distinguishability (see Methods). B, the divergence time among the studied species [44]. C, UMAP visualization of scVI integrated one-to-one orthologs only

data of gradually adding one more species among the five species. Colour showing species and cell types on the integrated embedding. D, self-projection test accuracy loss after scVI data integration, compared with unintegrated data per-species, in four integration tasks. With the addition of evolutionarily more distant species, cell types become blurred due to information loss during integration, result in a positive accuracy loss. E, UMAP visualization of the integration results for other four top-ranked integration strategies, coloured by species and cell types. SCCAF: single cell clustering assessment framework; H. sapiens: Homo sapiens; M. fascicularis: Macaca fascicularis; M. musculus: Mus musculus; X. laevis: Xenopus laevis; D. rerio: Danio rerio; MYA, million years ago; O2O, only use one-to-one orthologs; HE, one-to-one orthologs plus one-to-many and many-to-many orthologs matched by higher average expression level. AUC: area under curve; HM: human and macaque; HMM: human, macaque and mouse; HMMX: human, macaque, mouse and xenopus; HMMXD: human, macaque, mouse, xenopus and zebrafish.

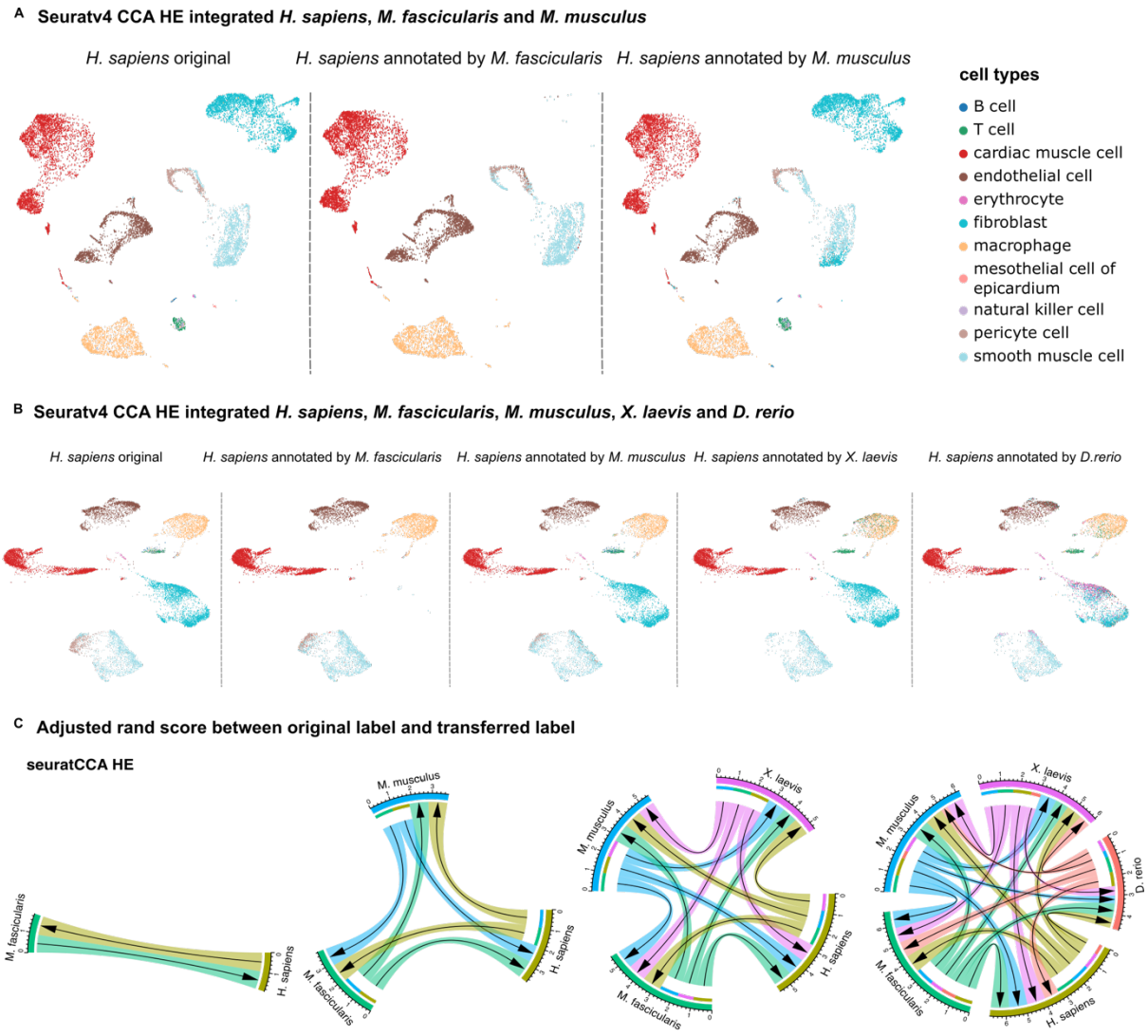


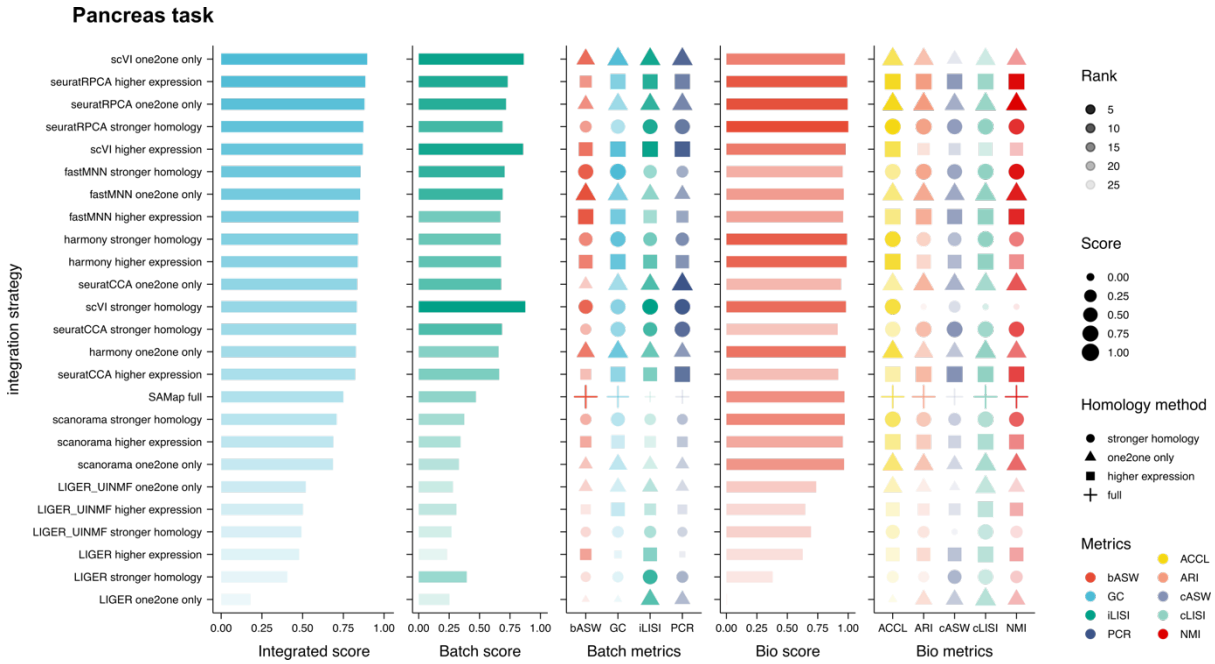
Figure 5 Cross-species cell type annotation transfer of heart tissue based on SeuratV4 CCA integrated data. A, UMAP visualization of annotation transfer using SeuratV4 CCA HE integrated human, long-tail macaque and mouse heart data. Colour shows the original cell type annotation of human data and the transferred annotation from macaque and mouse. Note that there are no cells annotated as 'fibroblast' in the macaque dataset. B, UMAP visualization of annotation transfer using SeuratV4 CCA HE integrated human, long-tail macaque, mouse, xenopus and zebrafish data. C, adjusted rand score of original annotation and transferred annotation for each pair of species during annotation transfer in four heart integration tasks. Annotation transfer is performed by applying Single Cell Clustering Assessment Framework (SCCAF) models between pairs of species. Each arrow starts with the species on which a SCCAF model for cell type classification is trained and ends with the species on which the model is applied to predict cell types. An ARI close to 1 indicates highly similar original and transferred cell type labels. H. sapiens: Homo sapiens; M. fascicularis: Macaca fascicularis; M. musculus: Mus musculus; X. laevis: Xenopus laevis; D. rerio: Danio rerio; HE: one-to-many and many-to-many orthologs matched by higher average expression level.

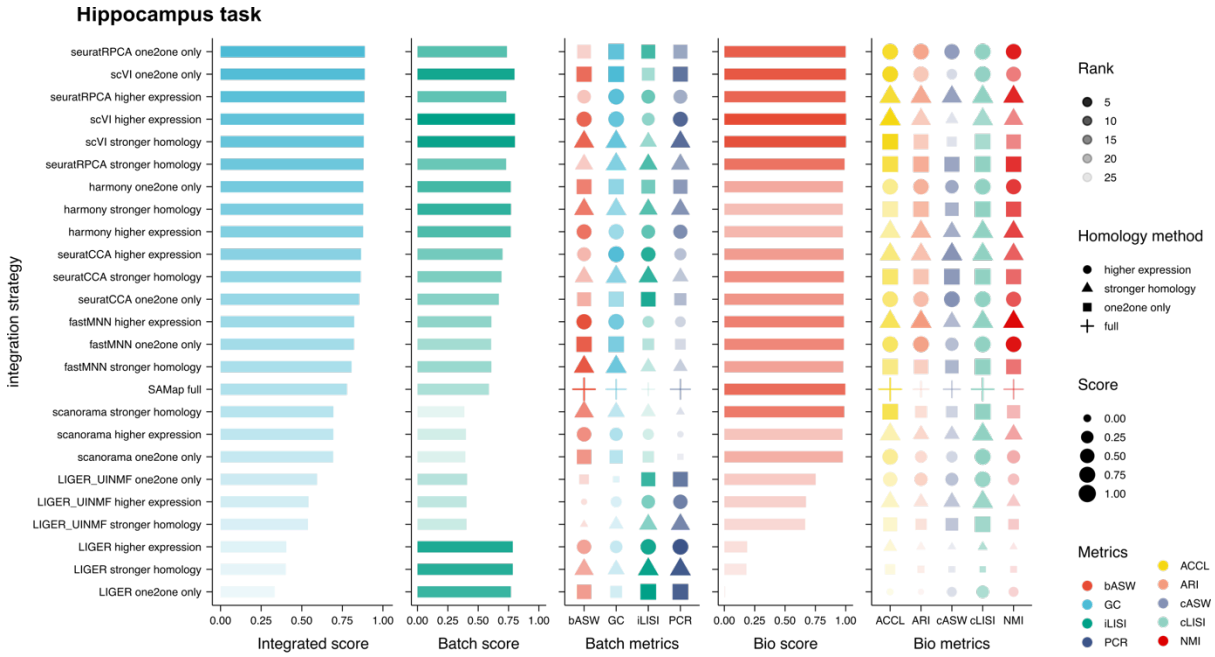
	Task name	Species	Technology	Number of cells	Number of homologous genes analysed		Assessment point	Data source reference
					one2one	one2many and many2many		
1	Pancreas_hs_mm	Homo sapiens, Mus musculus	inDrop	10,277	11,248	664	Basic performance	[27]
2	Hippocampus_hs_mm_ss	Homo sapiens, Macaca mulatta, Sus scroffa	snRNA-seq	116,788	12,557	24	Basic performance, large data size, complex subcluster structure	[28]
3	Embryo_xt_dr	Xenopus tropicalis, Danio rerio	inDrops	160,259	5,188	1,749	Whole-body data, developmental trajectory	[29,30]
4	Heart_hs_mf_mm	Homo sapiens, Macaca fascicularis, Mus musculus	scRNA-seq with 10× Genomics 3′ V3.1 droplet-based sequencing (H.sapiens), snRNA-seq (M.fascicularis), scRNA-seq	30,614	11,202	199	Cross-study and cross-technology integration	[1,3,31]

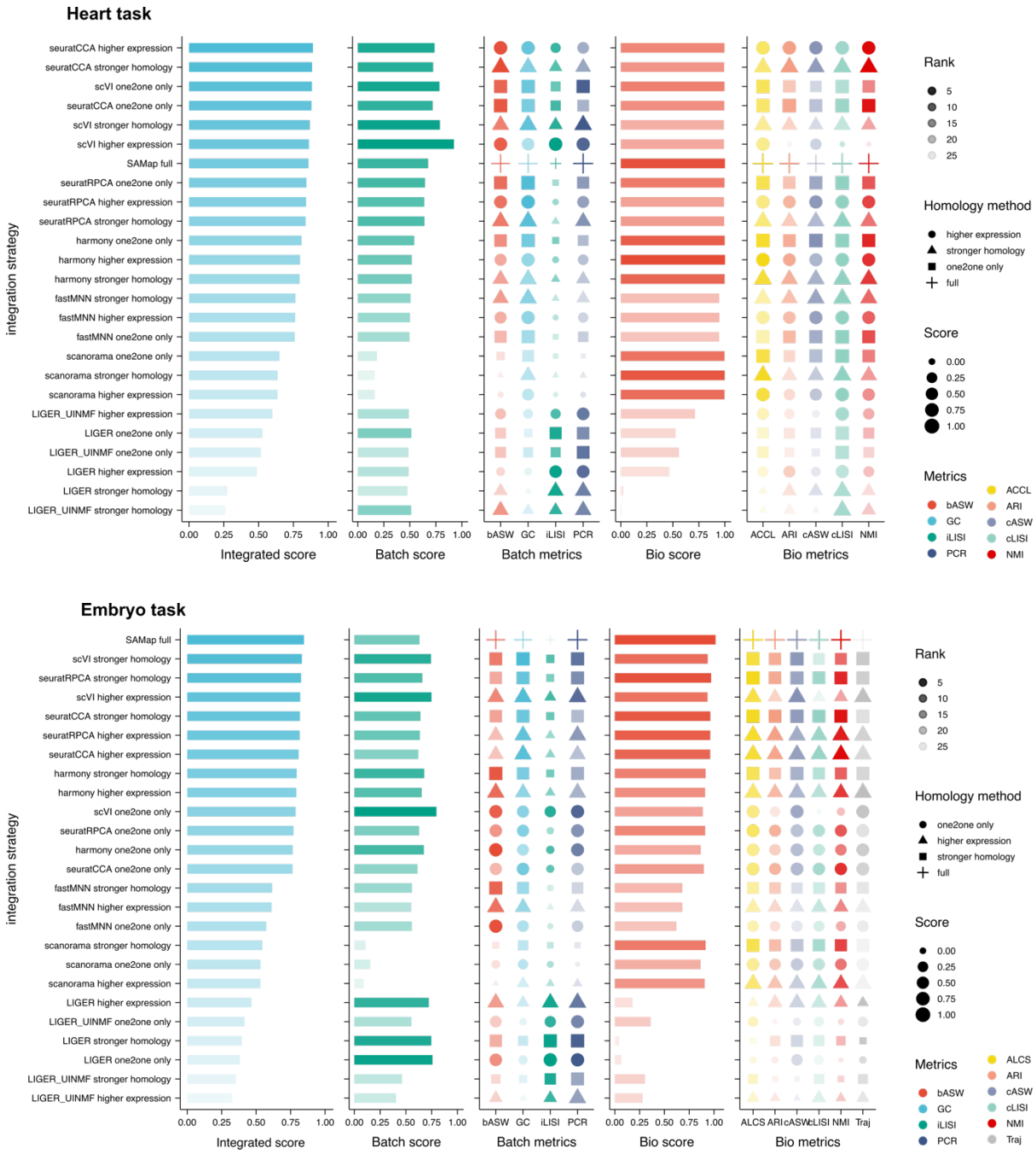
			(M.musculus)					
5	Heart_hs_mf	Homo sapiens, Macaca fascicularis	same with task 4	23,212	12,998	398	Close evolutionary distance	[1,3]
6	Heart_hs_mf_mm_xl	Homo sapiens, Macaca fascicularis, Mus musculus, Xenopus laevis	same with task 4, microwell-seq (X.laevis)	52,041	6,152	95	Far evolutionary distance, challenging homology annotation	[1,3,31,32]
7	Heart_hs_mf_mm_xl_dr	Homo sapiens, Macaca fascicularis, Mus musculus, Xenopus laevis, Danio rerio	same with task 6, microwell-seq (D.rerio)	55,234	3,695	81	Far evolutionary distance, challenging homology annotation	[1,3,5,31,32]

Table 1 Overview of the tasks used in this study for benchmarking cross-species integration strategies. Number of cells refers to cells passing quality control in the original literature that are included during the integration. Number of homologous genes refer to genes used during the integration according to following criteria: 1) quantified in each dataset 2) can be mapped to an ENSEMBL gene ID 3) have homology annotation among studies species in ENSEMBL multiple species comparison tool (version 106). One-to-one orthologs and in-paralogs, including one-to-many and many-to-many orthologs are defined in ENSEMBL and accessed via biomaRt. Assessment points indicates the challenges presented by the task. Tasks 1-4 are reference tasks used to score the integration strategies while tasks 4-7 are used to explore the maximum evolutionary distance between species until when data integration is still informative.

Extended data





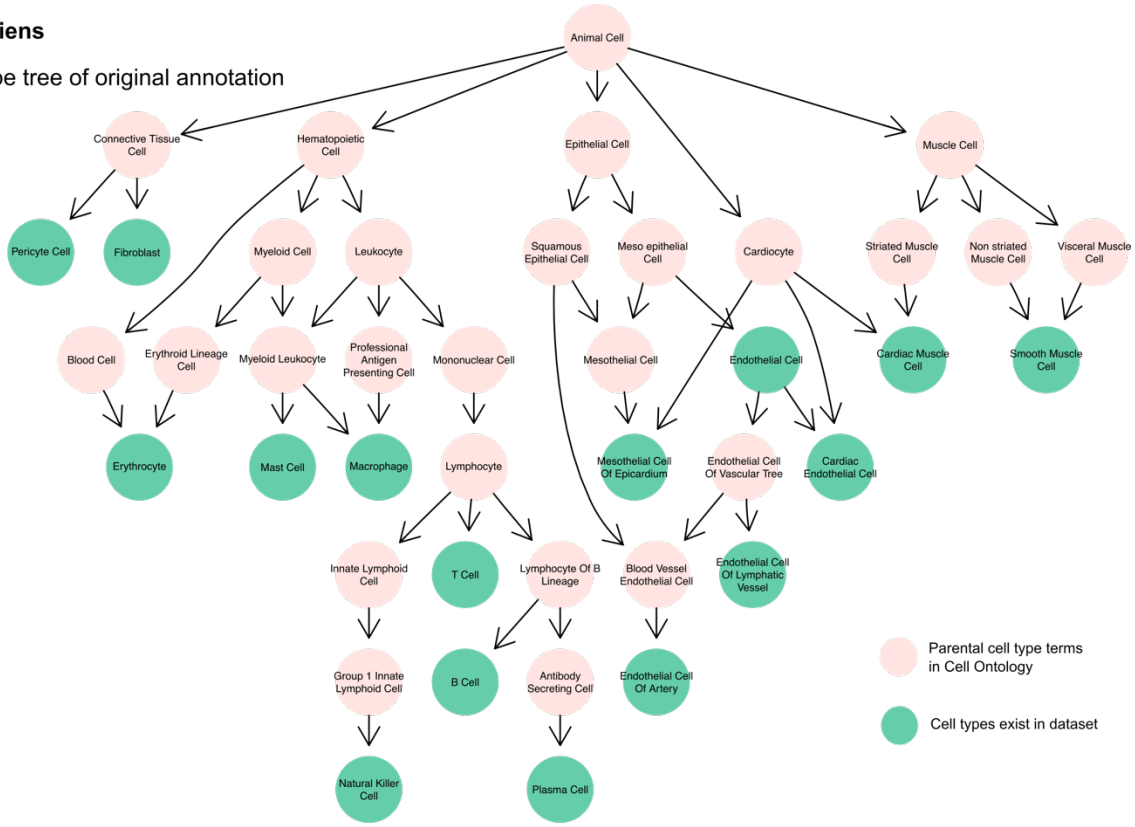


Extended Data Fig. 1-4: per-task benchmarking metrics and ranking of different integration strategies for four reference tasks. Integrated score is 0.4*species mixing score and 0.6*biology conservation score. Species mixing score is the average of 4 batch metrics, while biological conservation score is 0.75* ALCS plus 0.05* for the other 5 metrics. Batch metrics and biological conservation metrics are min-max scaled per task and the average scores across 4 tasks are shown in the figure. Trajectory conservation metric is only calculated for the relevant task: integrating embryonic development between xenopus and zebrafish. Bio: biology conservation; bASW: batch average silhouette width; GC: graph connectivity; iLISI: batch local inverse Simpson's index; PCR: principal component regression; ALCS: accuracy loss of cell type self-projection; ARI: adjusted rand index; cASW: cell

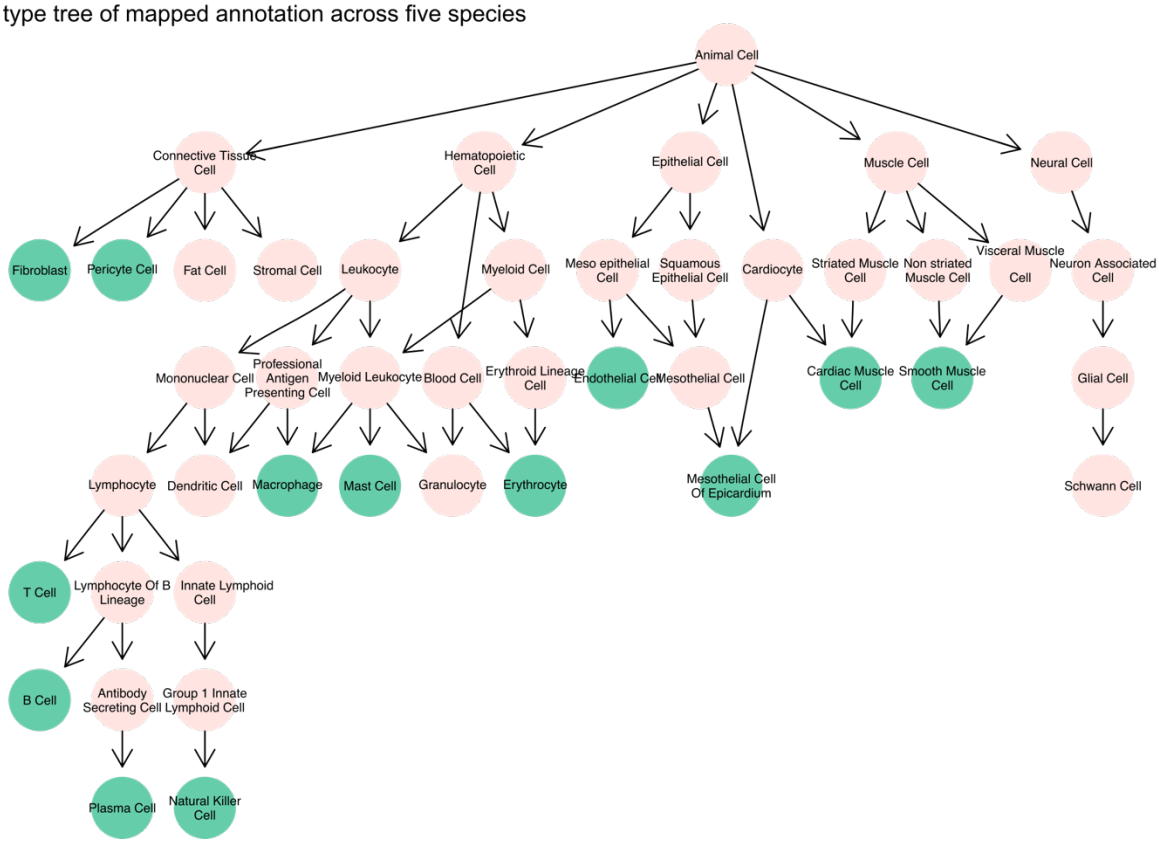
type average silhouette width; cLISI: cell type local inverse Simpson's index; NMI: normalized mutual information of cell label; Traj: trajectory conservation.

H. sapiens

Cell type tree of original annotation

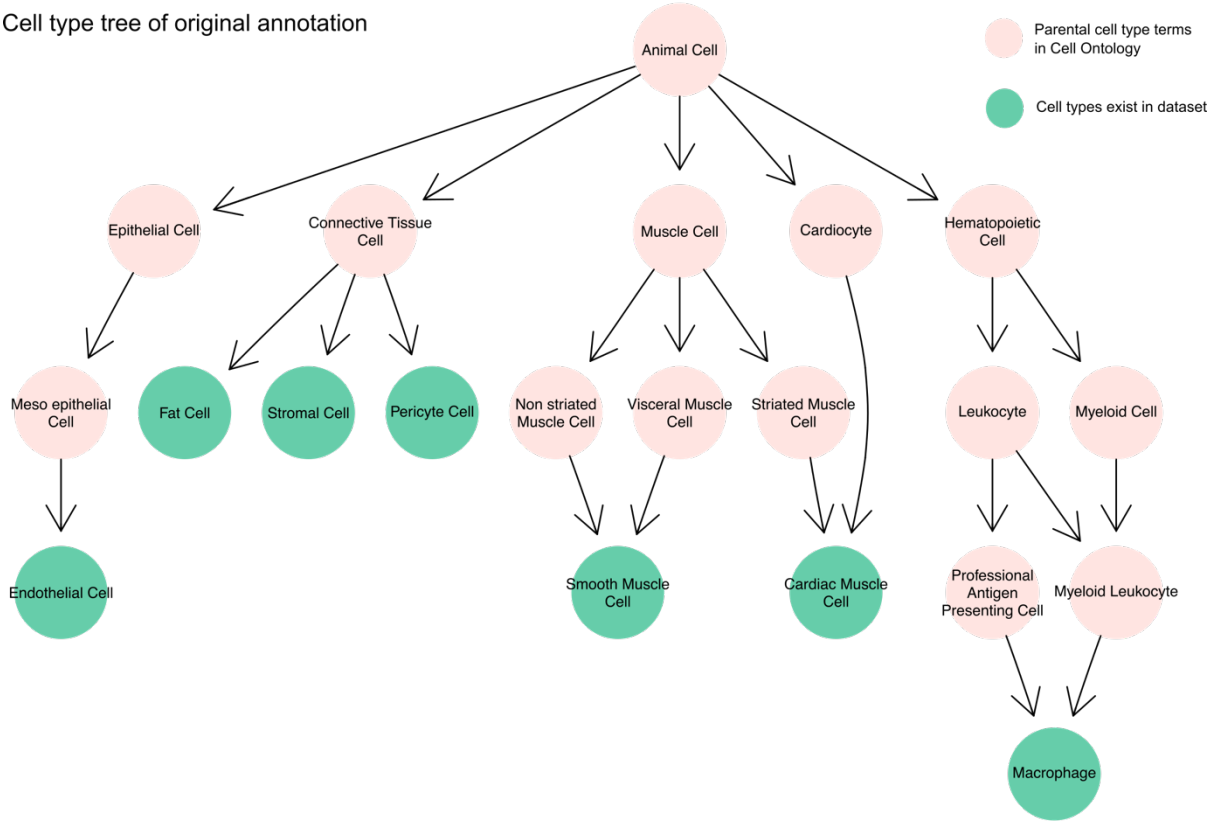


Cell type tree of mapped annotation across five species

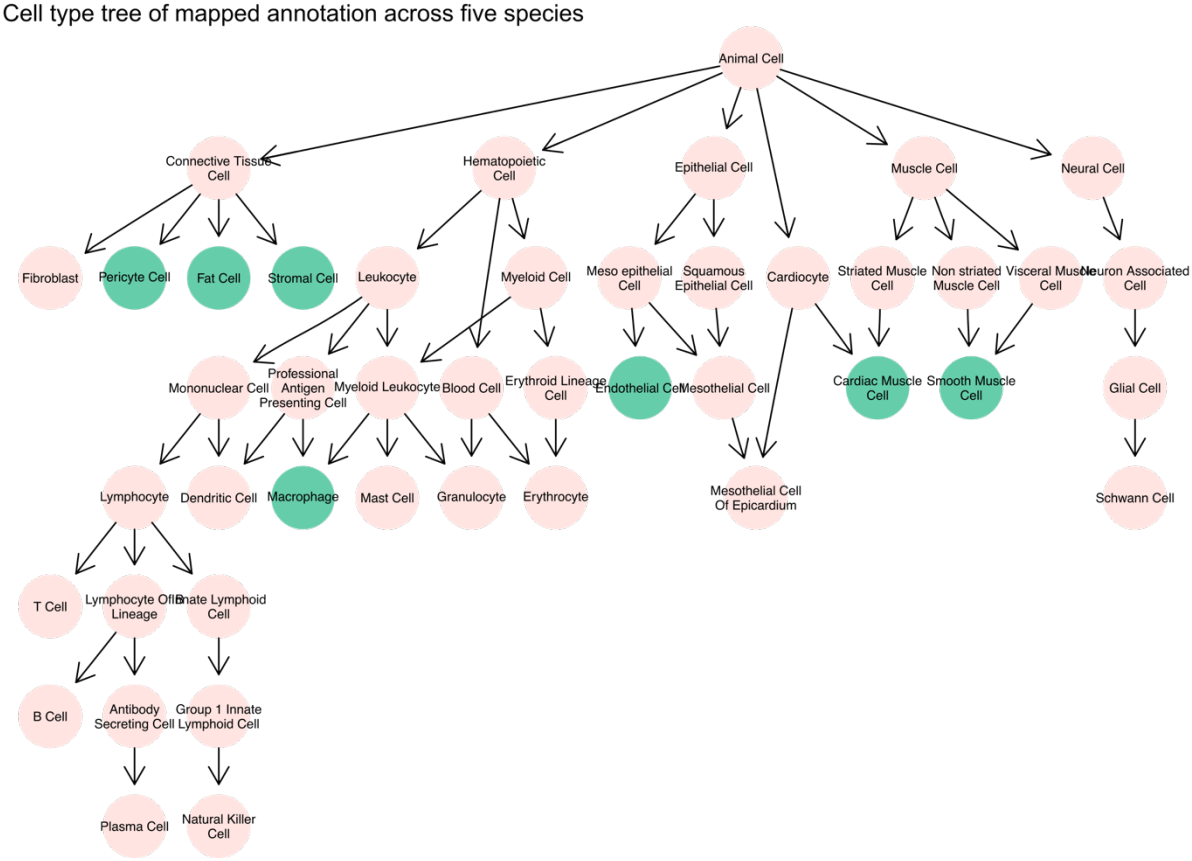


M. fascicularis

Cell type tree of original annotation

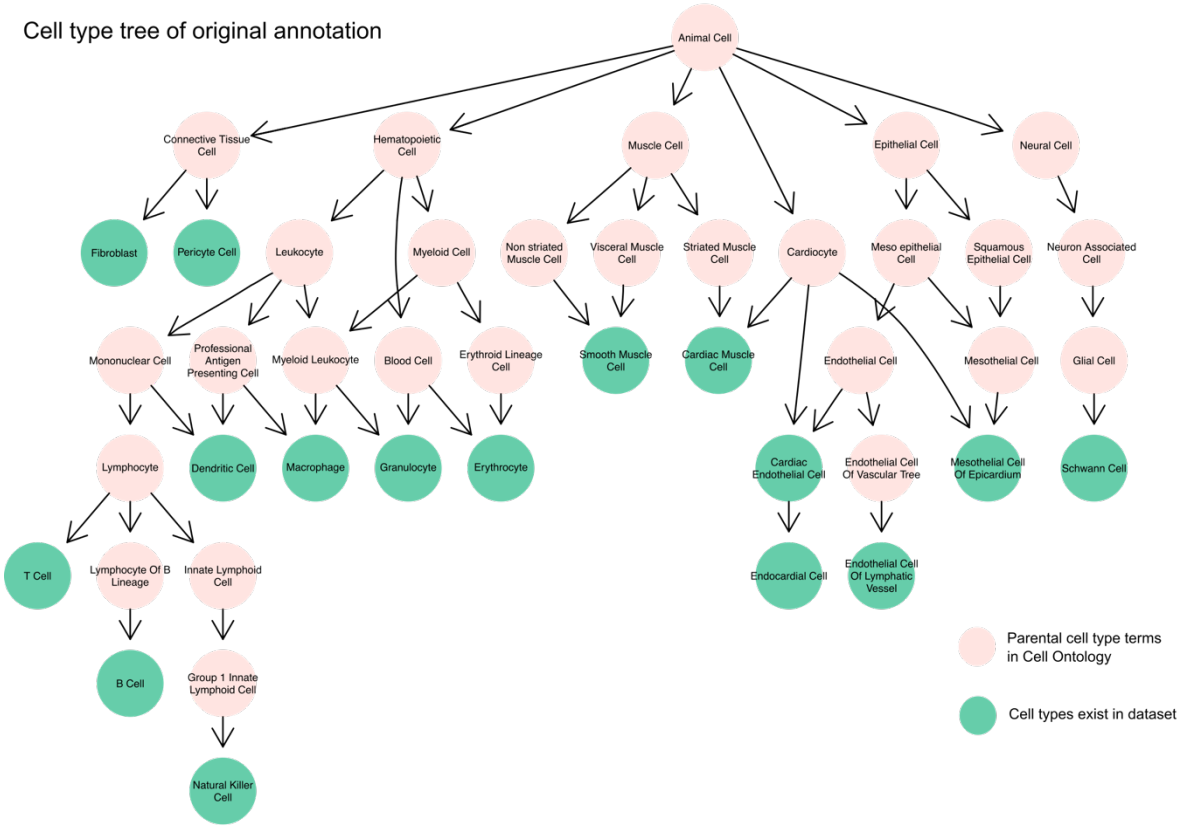


Cell type tree of mapped annotation across five species

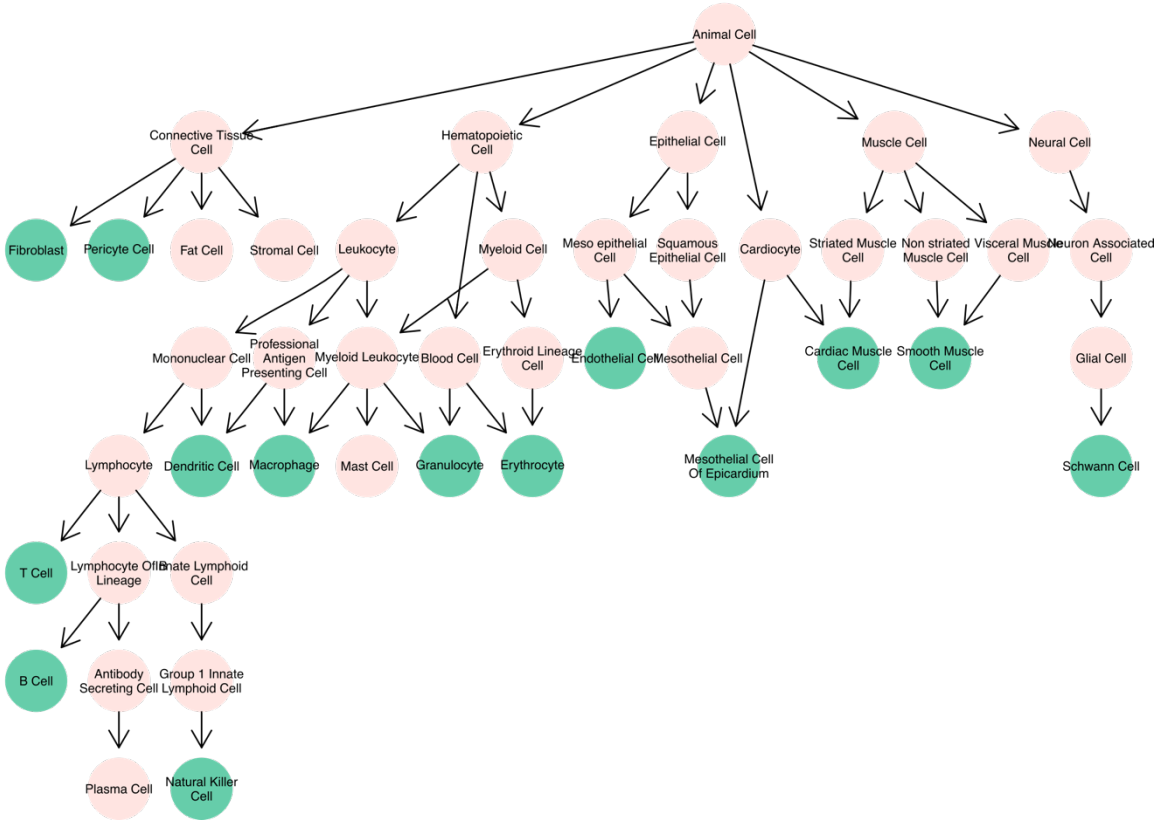


M. musculus

Cell type tree of original annotation

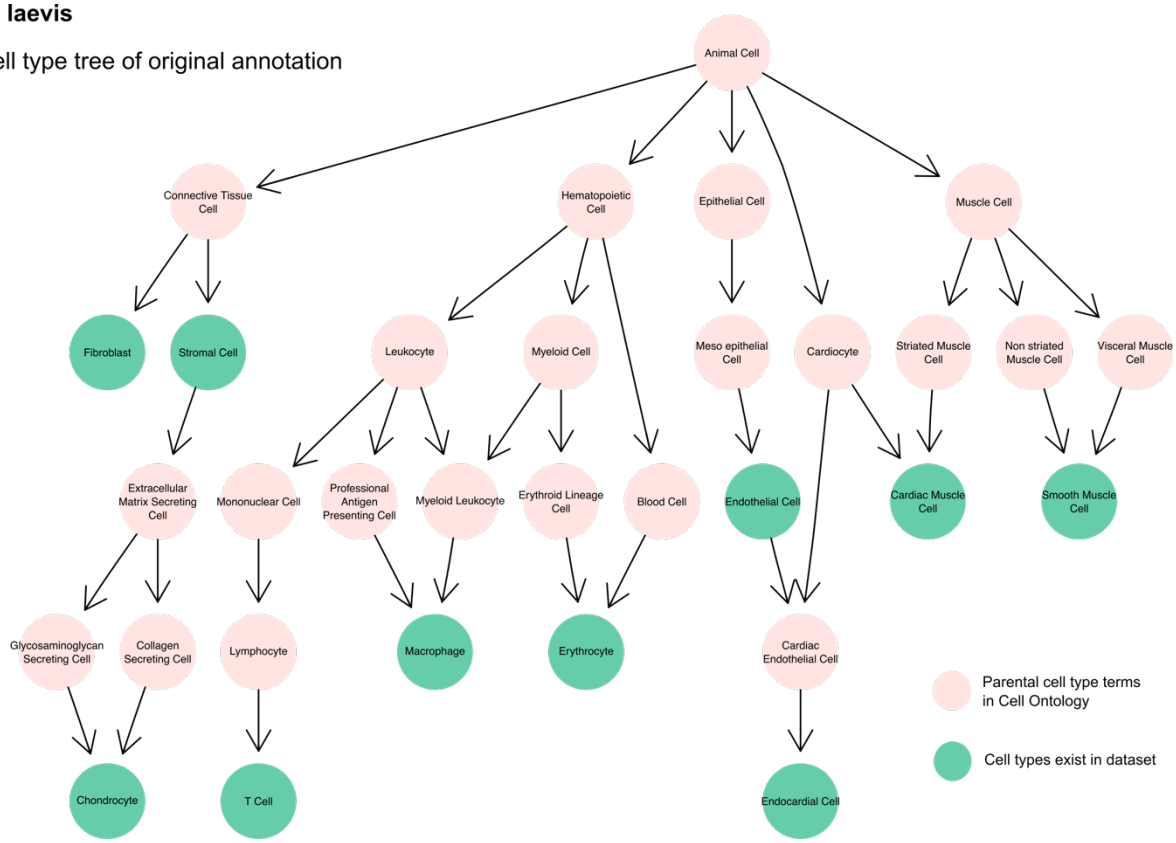


Cell type tree of mapped annotation across five species

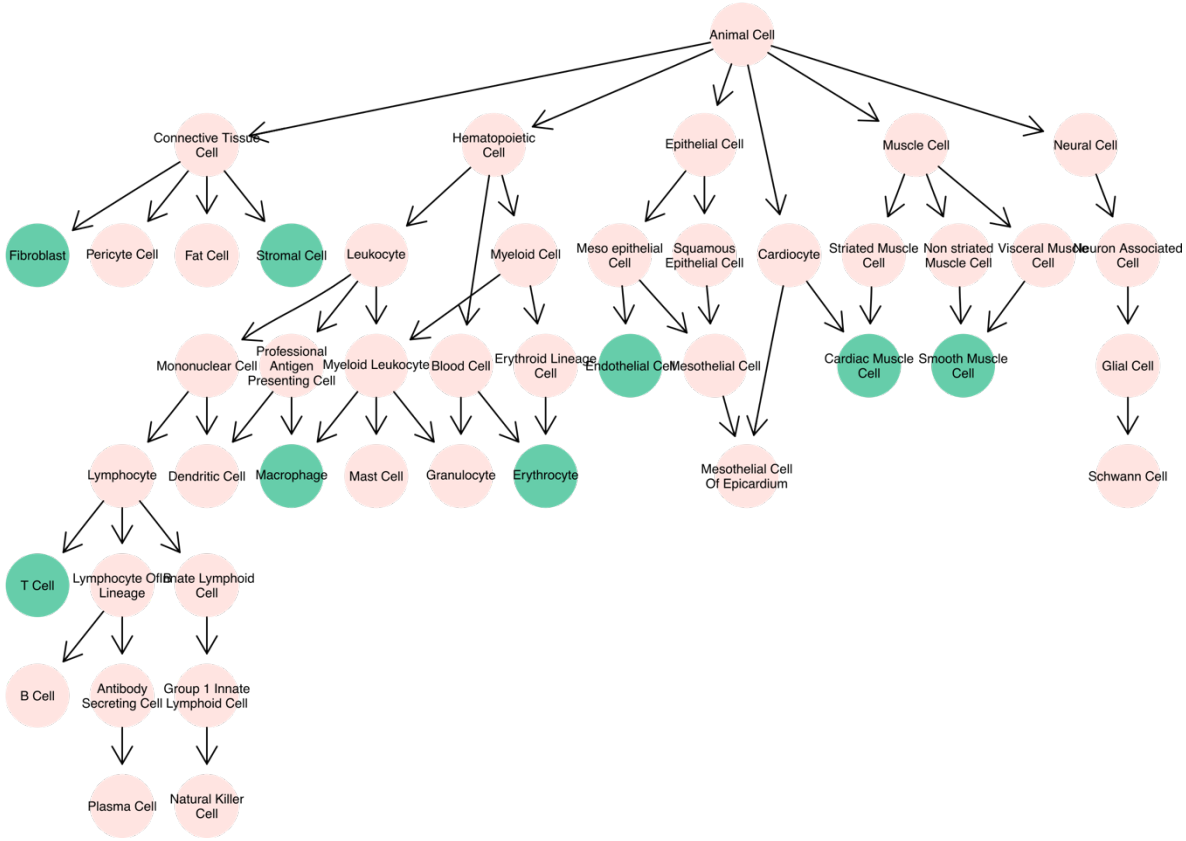


X. laevis

Cell type tree of original annotation

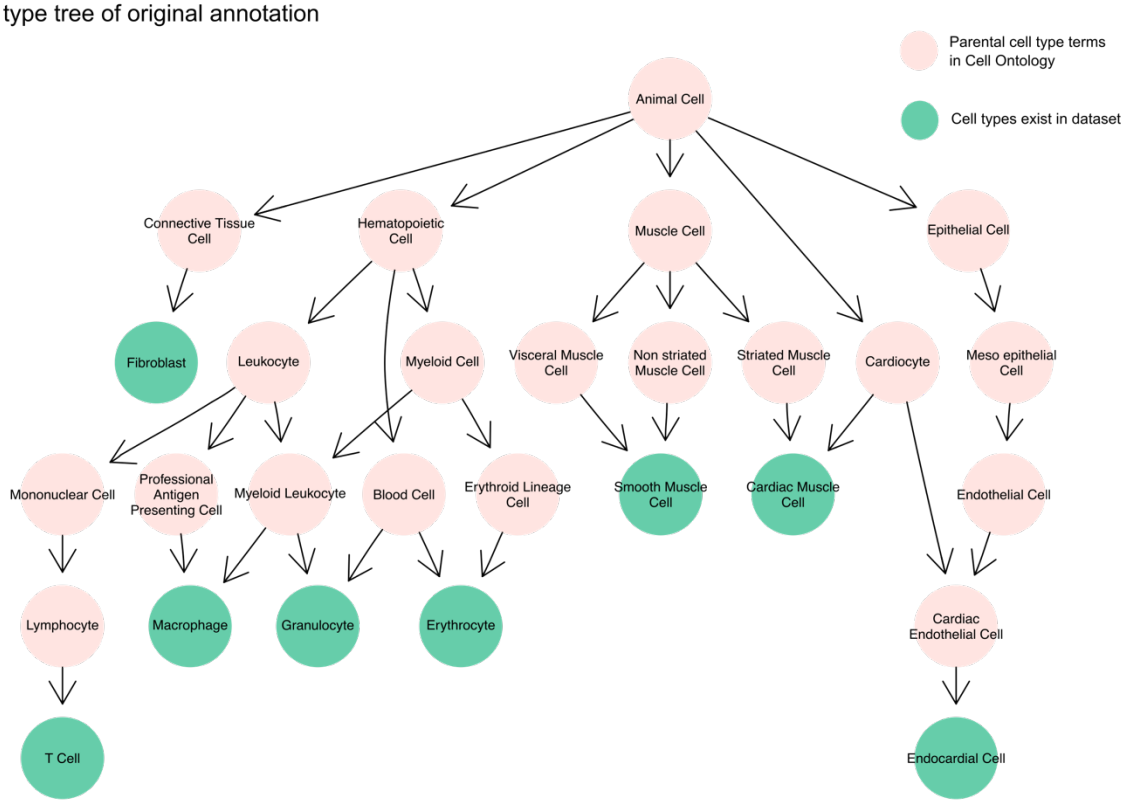


Cell type tree of mapped annotation across five species

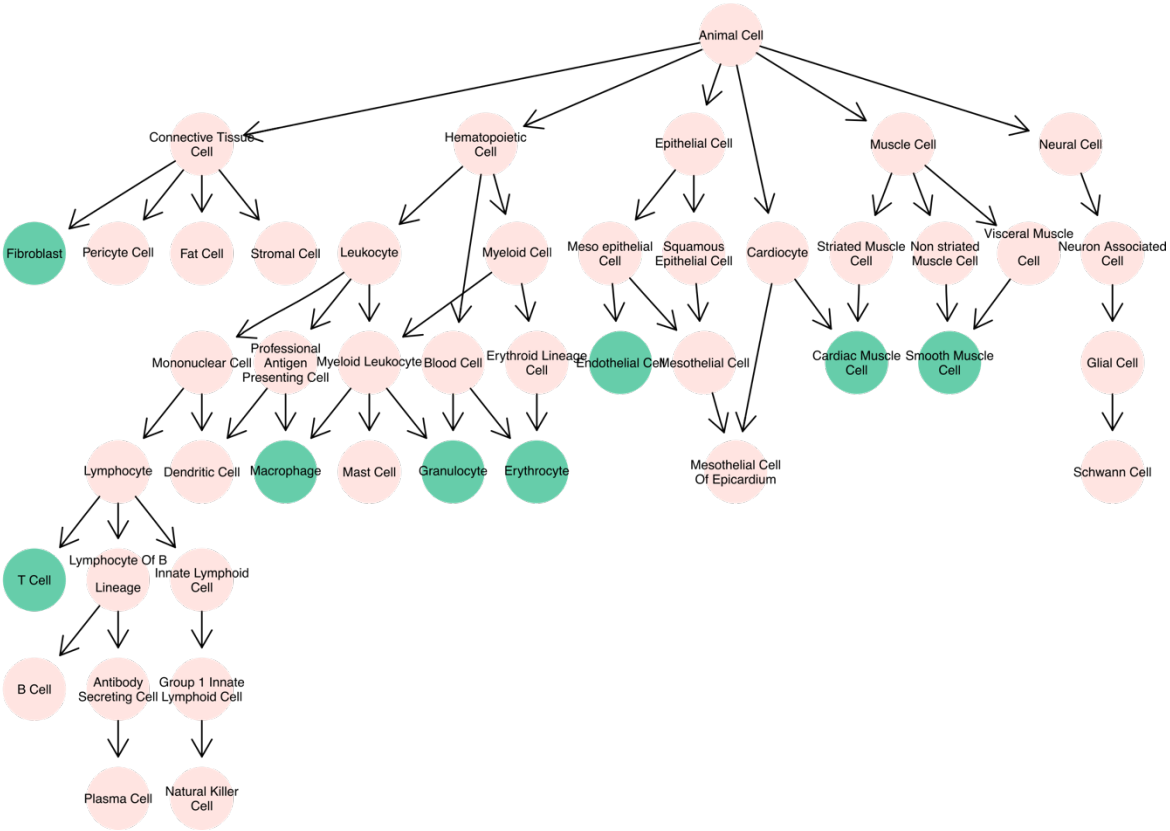


D. rerio

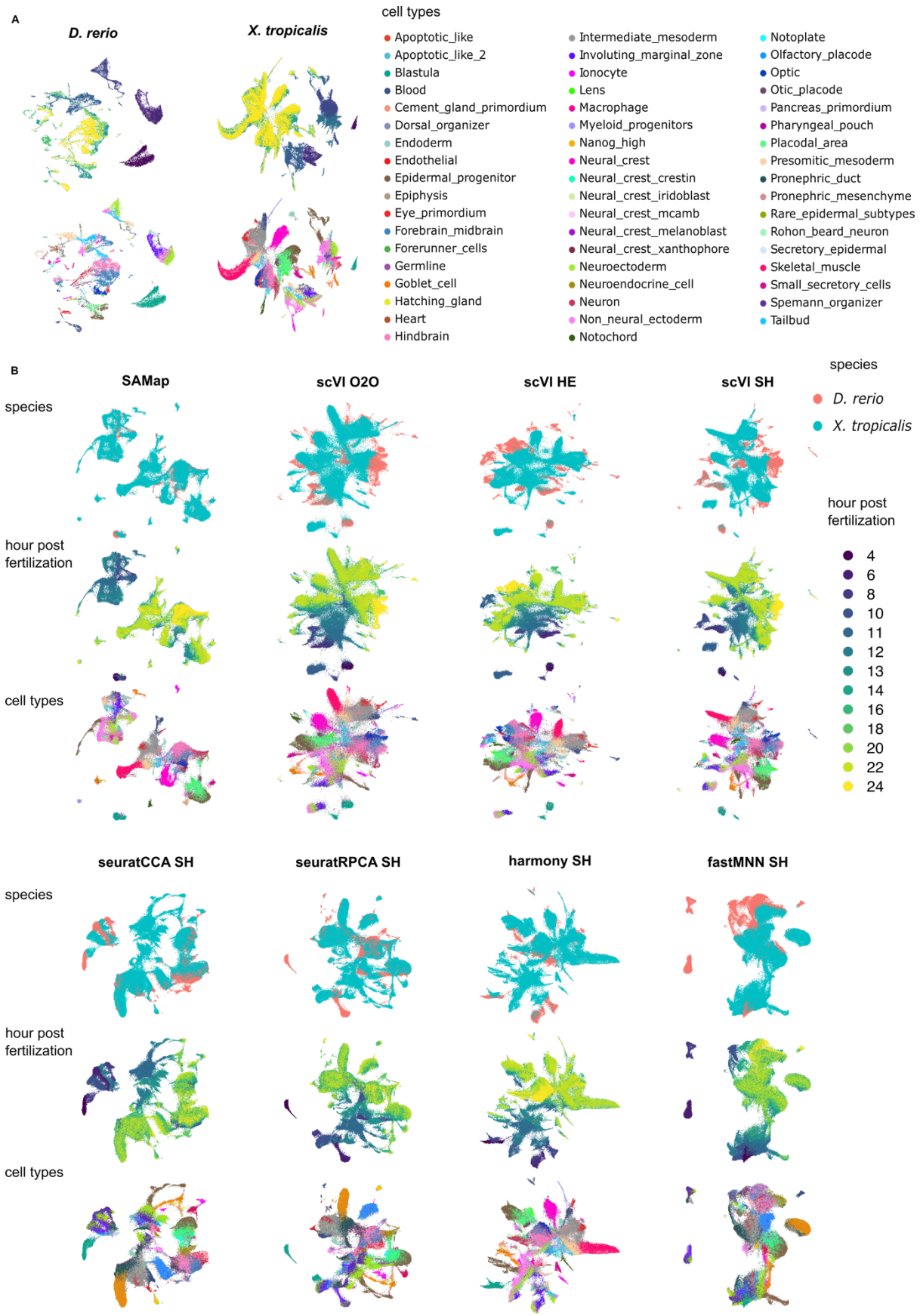
Cell type tree of original annotation



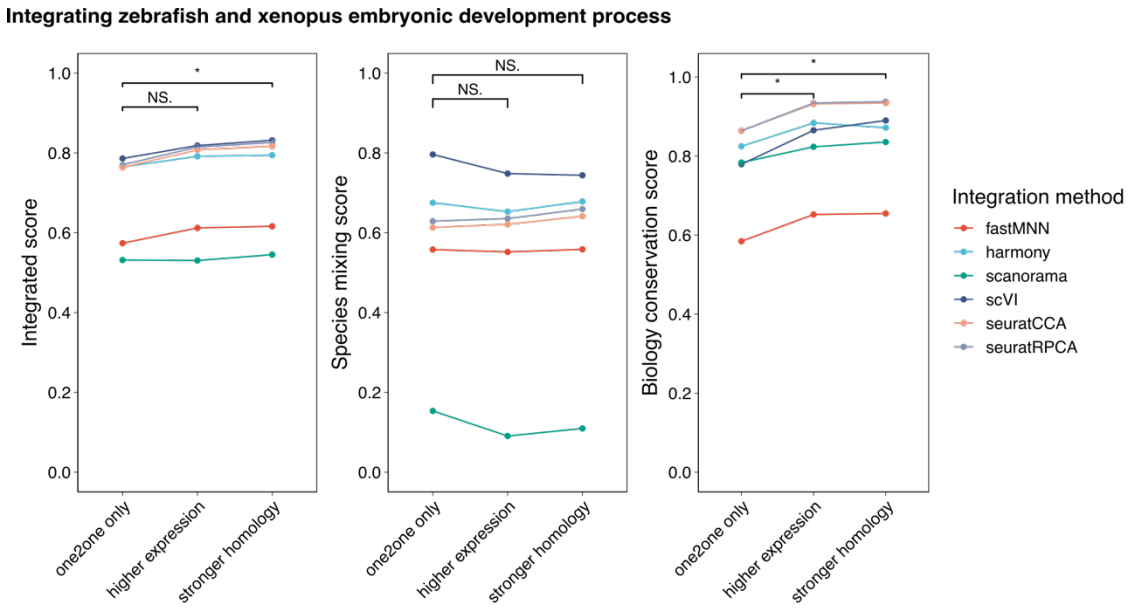
Cell type tree of mapped annotation across five species



Extended Data Fig. 5-9: scOntoMatch aligned cell type tree in heart task among five species. Each circle represents a cell type term in Cell Ontology system, coloured by whether they present in the dataset. Each arrow starts from a parental term and ends at its children term.


A
D. rerio
X. tropicalis
cell types
Apoptotic_like
Apoptotic_like_2
Blastula
Blood
Cement_gland_primordium
Dorsal_organizer
Endoderm
Endothelial
Epidermal_progenitor
Epiphysis
Eye_primordium
Forebrain_midbrain
Forerunner_cells
Germline
Goblet_cell
Hatching_gland
Heart
Hindbrain
Intermediate_mesoderm
Involuting_marginal_zone
Ionocyte
Lens
Macrophage
Myeloid_progenitors
Nanog_high
Neural_crest
Neural_crest_crestin
Neural_crest_iridoblast
Neural_crest_mcamb
Neural_crest_melanoblast
Neural_crest_xanthophore
Neuroectoderm
Neuroendocrine_cell
Neuron
Non_neural_ectoderm
Notochord
Notoplate
Olfactory_placode
Optic
Otic_placode
Pancreas_primordium
Pharyngeal_pouch
Placodal_area
Presomitic_mesoderm
Pronephric_duct
Pronephric_mesenchyme
Rare_epidermal_subtypes
Rohon_beard_neuron
Secretory_epidermal
Skeletal_muscle
Small_secretory_cells
Spemann_organizer
Tailbud
B
SAMap
scVI O2O
scVI HE
scVI SH
species
hour post fertilization
cell types
D. rerio
X. tropicalis
4
6
8
10
11
12
13
14
16
18
20
22
24
seuratCCA SH
seuratRPCA SH
harmony SH
fastMNN SH

Extended Data Fig. 10: embryonic development of zebrafish and xenopus integration results. A, UMAP visualization of cell types for zebrafish and xenopus embryo, coloured by cell types. B, UMAP visualization of integration results of various strategies in the embryo task. Coloured by species, hour post fertilization and cell type. O2O: only use one-to-one orthologs; HE: one-to-one orthologs plus one-to-many and many-to-many orthologs matched by higher average expression level; SH: one-to-one orthologs plus one-to-many and many-to-many orthologs matched by stronger homology confidence; D.rerio: Danio rerio; X.tropicalis, Xenopus tropicalis.



Extended Data Fig. 11: comparison of integrated score, species mixing score and biology conservation score between different homology methods in the embryo task. Paired sample Wilcoxon signed rank test were performed to compare between homology methods. NS.: non-significant; *:P-value < 0.05.