GigaScience

Improved integration of single cell transcriptome data demonstrates common and unique signatures of heart failure in mice and humans
--Manuscript Draft--

Manuscript Number:	GIGA-D-23-00222R2
Full Title:	Improved integration of single cell transcriptome data demonstrates common and unique signatures of heart failure in mice and humans
Article Type:	Research
Funding Information:	Deutsche Forschungsgemeinschaft (DFG) (Exc2026/1) — Dr. David John
	Dr. Rolf M. Schwiete Stiftung (Projekt 08/2018) — Prof. Dr. Stefanie Dimmeler

| Abstract: | Background |

Cardiovascular research heavily relies on mouse (mus musculus) models to study disease mechanisms and to test novel biomarkers and medications. Yet, applying these results to patients remains a major challenge and often results in non-effective drugs. Therefore, it is an open challenge of translational science to develop models with high similarities and predictive value. This requires a comparison of disease models in mice with diseased tissue derived from humans.

Results

To compare the transcriptional signatures at single cell resolution, we implemented an integration pipeline called OrthoIntegrate which uniquely assigns orthologs and therewith merges single cell data (scRNA-SEQ) of different species. The pipeline has been designed to be as easy to use and is fully integrable in the standard Seurat workflow.
We applied OrthoIntegrate on scRNA-SEQ from cardiac tissue of heart failure patients with reduced ejection fraction (HFrEF) and scRNA-SEQ from the mice after chronic infarction, which is a commonly used mouse model to mimic HFrEF. We discovered shared and distinct regulatory pathways between human HFrEF patients and the corresponding mouse model. Overall, 54% of genes were commonly regulated including major changes in cardiomyocyte energy metabolism. However, several regulatory pathways, e.g. angiogenesis, were specifically regulated in humans.

Conclusion

The demonstration of unique pathways occurring in humans indicate limitations on the comparability between mice models and human HFrEF and show that results from the mice model should be validated carefully. OrthoIntegrate is publicly accessible (https://github.com/MarianoRuzJurado/OrthoIntegrate) and can be used to integrate other large data sets to provide a general comparison of models with patients data.

Corresponding Author:	David John, Ph-D Goethe-Universitat Frankfurt am Main Frankfurt am Main, Hessen GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Goethe-Universitat Frankfurt am Main
Corresponding Author's Secondary Institution:	
First Author:	Mariano Ruz Jurado
First Author Secondary Information:	
Order of Authors:	Mariano Ruz Jurado

	Lukas S. Tombor
	Mani Arsalan
	Tomas Holubec
	Fabian Emrich
	Thomas Walther
	Wesley Abplanalp
	Ariane Fischer
	Andreas M. Zeiher
	Marcel H. Schulz
	Stefanie Dimmeler
	David John, Ph-D
Order of Authors Secondary Information:	
Response to Reviewers:	Response to the reviewers:

Again, we would like to thank the reviewers for the remarks and valuable suggestions. And also for the fair evaluation of our scientific work. We addressed all the point in the section below and hope that the paper is then acceptable for publication.

Reviewer 2:
The author's additional analysis is commendable. With the inclusion of new evaluation metrics, the benchmark section now appears relatively comprehensive, and the explanations provided for the reduced NMI score are reasonable. In the results section, the supplementary information on functional enrichment further elucidates the biological functions of fibroblast cluster 25 and endothelial cell cluster 28. There are still some minor suggestions for improvement:

1. The presentation of the biological findings in the discussion section could be more succinct to improve clarity.

We agree with the reviewer and tried to shorten the biological discussion section of the manuscript as much as possible. However, in order to show the advantage of OrthoIntegrate to perform a single cell side-by-side comparison of from different species, we think that it is crucial to at least discuss one example for human or mouse specific pathways/genes and commonly regulated genes. Thereby we can exemplify the usability of OrthoIntegrate and show potential research targets for other researchers.

2. There is a lack of discussion on the impact of the numerous lncRNAs generated by OrthoIntegrate. This topic requires further exploration and elaboration.

According to the suggestions we added the following paragraph, to the discussion section of the paper:

"Due to the increased numbers of features that are included in OrthoIntegrate, the clustering might be more diverged, likely by species specific non-coding RNAs or other features, which are not included in the other databases. Therefore, the more divergent clustering, due the increased number of features in OrthoIntegrate combined with the broad cell type labeling might explain the slightly reduced NMI scores. However, since various publications have shown that long-non-coding RNAs have important regulatory roles in the heart [42–44], we think that these additional non-coding RNA's are an important resource to study species specific responses to different disease condition, especially in the field of heart failure."

3. Reorganize the paragraphs for "Single cell pre-processing" and "Study samples" to clarify the source of the data used in the article. Emphasize the data generated by authors (E-MTAB-13264) and provide details on the single-cell sequencing process

	(not only the raw data pre-processing).
	We thank the reviewer for this remark. Accordingly we reorganized the paragraphs and added the citation which contains the exact protocol on how the nuclei isolation was done and how RNA was processed and sequenced. We added the following paragraph to the paper: "Nuclear isolation steps and single-nucleus RNA-sequencing library preparation were conducted as described in Nicin et al.; NCVR 2022 (Nicin et al. 2022). "

Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be	Yes

either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript.

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?

Improved integration of single cell transcriptome data demonstrates common and unique signatures of heart failure in mice and humans

Mariano Ruz Jurado[,1,2,3], Lukas S. Tombor[1,2], Mani Arsalan[4], Tomas Holubec[4], Fabian Emrich[4], Thomas Walther[2,3,4], Wesley Abplanalp[1,2,3], Ariane Fischer[1], Andreas M. Zeiher [1,2,3], Marcel H. Schulz[1,2,3], Stefanie Dimmeler[1,2,3], David John*[1,3]

1.) Institute of Cardiovascular Regeneration, Theodor-Stern-Kai 7, 60590 Frankfurt am Main

2.) German Centre for Cardiovascular Research (DZHK), Frankfurt am Main, Germany

3.) Cardio-Pulmonary Institute (CPI), funded by the German Research Foundation (DFG)

4.) Department of Cardiovascular Surgery, Goethe University Hospital, 60590 Frankfurt am Main

*Corresponding author: David John, email: john@med.uni-frankfurt.de

ORCID iDs:

Mariano Ruz Jurado [0000-0001-5354-5336]; Lukas S. Tombor [0000-0002-5867-576X]; Mani Arsalan [0000-0001-5861-891X]; Tomas Holubec [0000-0003-2366-0880]; Fabian Emrich [0000-0003-4546-889X]; Thomas Walther [0000-0002-6725-4473]; Wesley Abplanalp [0000-0001-5996-8997]; Ariane Fischer [0000-0001-8074-3354]; Andreas M Zeiher [0000-0003-1711-5819]; Marcel H Schulz [0000-0002-1252-3656]; Stefanie Dimmeler [0000-0002-1045-2436]; David John [0000-0003-3217-5449]

Abstract

BackgroundOperational Delay

Cardiovascular research heavily relies on mouse (*mus musculus*) models to study disease mechanisms and to test novel biomarkers and medications. Yet, applying these results to patients remains a major challenge and often results in non-effective drugs. Therefore, it is an open challenge of translational science to develop models with high similarities and predictive value. This requires a comparison of disease models in mice with diseased tissue derived from humans.

Results

To compare the transcriptional signatures at single cell resolution, we implemented an integration pipeline called *OrthoIntegrate* which uniquely assigns orthologues and therewith merges single cell data (scRNA-SEQ) of different species. The pipeline has been designed to be as easy to use and is fully integrable in the standard Seurat workflow.

We applied *OrthoIntegrate* on scRNA-SEQ from cardiac tissue of heart failure patients with reduced ejection fraction (HFrEF) and scRNA-SEQ from the mice after chronic infarction, which is a commonly used mouse model to mimic HFrEF. We discovered shared and distinct regulatory pathways between human HFrEF patients and the corresponding mouse model. Overall, 54% of genes were commonly regulated including major changes in cardiomyocyte energy metabolism. However, several regulatory pathways, e.g. angiogenesis, were specifically regulated in humans.

Conclusion

The demonstration of unique pathways occurring in humans indicate limitations on the comparability between mice models and human HFrEF and show that results from the mice model should be validated carefully. *OrthoIntegrate* is publicly accessible (https://github.com/MarianoRuzJurado/OrthoIntegrate) and can be used to integrate other large data sets to provide a general comparison of models with patients data.

Keywords

cross-species analysis, cardiovascular disease, heart failure with reduced ejection fraction, coronary artery ligation, single cell integration, cross species integration workflow

Introduction

Animal experiments are a powerful tool to improve our understanding of pathophysiological conditions and to predict responses to new therapeutic approaches [1]. However, due to ethical considerations they are controversially discussed [2], and their predictive capacity for toxicity and drug responses is limited [3,4]. Especially mice are commonly used to model human diseases as they are relatively inexpensive, have short generation times and have large numbers of offspring. Additionally mice have a relatively close physiological and phylogenetic relationship with humans [5], [6]. Mice protein-coding genes are on average 85% identical to humans [4] and over 90% of both genomes have regional conserved synteny [7]. Due to these advantageous breeding characteristics and their high sequencing conservation to humans, hundreds of different mouse models have been developed to study human diseases [8] like heart failure [9] or even diseases that do not occur naturally in mice like Alzheimer's or Parkinson's disease [10].

To study cardiovascular diseases, which remain the leading cause of morbidity and mortality in the aging society, the ligation of the left anterior descending coronary artery model (LAD) is often used to induce myocardial infarction, which results in ischemic heart failure with reduced ejection fraction (HFrEF) [11,12]. Thereby, the LAD is ligated to mimic the clotted artery as it occurs after infarction. While short term reperfusion then allows to mimic the reopening of the coronary artery by catheter based interventions, often chronic ligation is used to induce heart failure over the course of > 4 weeks. As this method describes a similar decline in heart function, scientists use LAD mouse models to simulate HFrEF and develop and test new therapeutic strategies [13–15]. Patients who suffer from HFrEF, are unable to pump sufficient amounts of blood to meet the demands of body organs [16].

To address the comparability of HFrEF in human to mouse models, we used single nuclei RNA sequencing data, enabling us to assess transcriptional regulatory pathways in all cardiac cell populations with high resolution and accuracy [17,18]. In order to analyze scRNA-seq data from various samples, integration pipelines were developed to combine individual cells from

different subjects into clusters with similar expression patterns [18,19]. Yet these bioinformatic tools can only integrate datasets from identical species. Several studies developed algorithms to compare mRNA expression patterns across species [20–22]. However, a standardized and easy way to compare single cell/nuclei RNA sequencing data sets of human and mouse by directly integrating the data is still missing [18,23,24]. To overcome these limitations and the highly increasing demand for comparison of various organisms prompted us to develop a R package called *OrthoIntegrate*. It features a pipeline for integration of single cell datasets and orthologue assignment, allowing the simple integration of data from animal models and human patients. For the orthologue assignment process, we implemented an algorithm in the workflow that adjusts the different nomenclature between species before the integration takes place, by using the databases of Ensembl, NCBI, and Uniprot. [25–27]. Using our newly established pipeline, which is completely compatible with standard seurat workflows, we explored the gene expression patterns in mouse models of HFrEF compared to human samples. While 54% of genes were commonly regulated in both species, we also observed significant differences in differentially expressed genes and regulated pathways in patients with heart diseases than in the corresponding mouse model.

Results

One to one orthologue assignments

To integrate single cell data from different species, we established a table of gene names, which contains one human gene for each mouse gene, by which it will be replaced (one-to-one orthologues). We performed the same approach for generating a table of gene names between human and zebrafish genes.

In order to generate these one-to-one orthologues, we utilized the Needleman-Wunsch algorithm [28] to perform a pairwise global alignment between possible orthologues retrieved by Ensembl database. This calculation determines alignment scores based on differences in the amino acid or nucleotide sequences. In case no orthologues were found, nor a protein- or

nucleotide sequence is available for a particular gene, a lowercase matching of the human gene is searched for in the mouse gene database (Suppl. Fig. 1A).

The Ensembl database assigned a total of 21,428 mouse orthologues to our human gene ID symbols. However, only 77% (16,573) of these were uniquely assigned. Through our *OrthoIntegrate* pipeline, we increased the number of assignments to 82% (17,504). Hereby, 714 mouse genes were assigned by protein sequence alignment, 89 through nucleotide sequence alignment, 42 by using the Levenshtein distance between gene names and 86 using our lowercase matching approach. Most of the 86 matches found by lowercasing were long non coding RNAs with identical gene names. We then proceeded by filtering the human and mice data by these orthologues in our pipeline and replaced the mice nomenclature by the human nomenclature for the corresponding samples (Suppl. Fig. 1B). In the end, we could assign ~82% of the mice genes to human orthologues (Suppl. Table 2). Replacing mouse gene names with the human orthologue allowed us to integrate the human patient data with the mouse model data into one single cell object (Fig. 1A). Moreover, we aimed to underscore the versatility of OrthoIntegrate. Therefore, we integrated and clustered scRNA-seq datasets related to Alzheimer's disease from human, mouse and zebrafish with *OrthoIntegrate* pipeline (Suppl. Fig. 6). We successfully created clusters representing excitatory and inhibitory neurons, as well as astrocytes, in the three species (Suppl. Fig. 6A-B). Given the focus and the size of the human study and data most of the excitatory neurons found were of human origin, but we showed that excitatory neurons found in mice were also assigned to the same clusters and showed comparable marker genes (Suppl. Fig. 6A-D). Similar results were obtained for inhibitory neurons and astrocytes proving a successful integration of all three datasets (Suppl. Fig. 6C-F).

Cell type composition in human and mouse upon HFrEF

After demonstrating the practicality of the integrated dataset, the biological differences of the human mouse datasets were analyzed. The absence of species specific clusters in the combined UMAP plot confirms that human and mouse hearts comprise similar cell types and

gene expression patterns (Fig. 1B). This is additionally verified by similar cell type specific marker genes in both species in the different cell clusters (Fig. 1E-F). The specific marker genes allowed the annotation of the clusters into cardiomyocytes (CMs), pericytes (PCs), smooth muscle cells (SMCs), fibroblasts (FBs), endothelial cells (ECs), immune cells (ICs), as well as neuronal cells (NCs) (Fig. 1C). In addition, we analyzed how the distribution of cell types was affected by the heart failure phenotype. Thereby a 20% decrease in human CM was observed when comparing the control samples with the HFrEF samples (45% -> 25%) (Fig. 1D). However, in mice, there is no difference in the numbers of CMs between the infarcted and control mice (both comprise about ~25% CM) (Fig. 1D). Furthermore, we found differences in the distribution of ECs in the human versus mouse samples. Specifically, we observed a significant increase in the EC population in samples from HFrEF patients (~30 %) compared to healthy hearts (~8%). In contrast, we noticed decreased EC numbers in mice upon infarction (from 25% in controls to 18% after chronic infarction). Minor changes are also observed in the contributions of other cell types (Fig. 1D).

Comparison to other integration methods

We carefully inspected our data to determine species specific distribution by creating UMAP plots of all cells in our integrated object. Figure 1B shows that cells of mouse and human origin commingled in all clusters, which indicates a successful integration based on the cell types and not on the species. We additionally compared our *OrthoIntegrate* pipeline to other orthologue databases and tools to assess the advantages of our orthologue assignments. For this purpose, we created the same scRNA-SEQ datasets using the different orthologue lists OMA, Biomart and InParanoid [29–31]. Visualization of the integration by UMAP plots shows an integration of human and mouse-derived cells in the individual cell clusters also with the alternative orthologous list (Suppl. Fig. 2A-C). However, besides the visual impression, quantitative metrics were used to assess the quality of the clustering, we calculated silhouette coefficients, which measure the quality of the clustering independent from the number of

clusters. Integration by *OrthoIntegrate* resulted in the highest silhouette coefficients compared to the other orthologue databases, suggesting an improved clustering (Fig. 2A). Additionally, it is noteworthy to mention that our pipeline achieved by far the most 1:1 protein coding and lncRNA coding orthologous pairs in comparison to the other described methods (Fig. 2B-C). To further determine the clustering quality after integration, we computed supplementary metrics recommended by the single-cell integration benchmark scib package [32] and the Orthology Benchmark Service. We also calculated the Species Mixing Score and Bio Conservation Score, following the guidelines of the BENGAL pipeline (Fig. 2D) [32,33]. Remarkably, our method not only achieved the highest number of uniquely mapped orthologous pairs but also demonstrated high performance across individual metrics in comparison with alternative tools (Fig. 2D-F).

Differential gene expression between mice and humans

The differentially expressed gene (DEG) analysis, showed strong similarities in the regulated genes upon HFrEF. However, some genes showed differences in their expression patterns. Mainly when the cell types were analyzed individually. Overall, we found a comparable number of DEGs in both species (4141 in humans, 4654 in mice).

The average of commonly regulated genes per cell type (Fig 3A; left side) showed that around 54% of DEGs found in humans were also regulated in mice, with minor differences between cell types. Upregulated genes showed a generally higher comparability compared to downregulated genes (Fig. 3B). Only in smooth muscle cells many more human specific DEGs were regulated in opposite directions (Fig. 3B, right upper panel). Averaging the mouse regulated DEGs (Fig 3A; right side) showed that only about 34% of the cell type specific DEGs in mice were regulated in humans, indicating a more substantial transcriptional effect of the LAD model compared to the human disease.

Figure 3B separately shows the up (top panel) and down (lower panel) regulated genes in humans and their regulation in mice. For the upregulated genes in humans, around 50-70%

of the corresponding mouse genes were also upregulated, around 25% were not regulated and only about 5-20% percent were regulated in the opposite direction. Suggesting that overall activation occurs mainly in similar expression pathways across all cell types. In the downregulated genes in humans, we observed a strikingly low number of commonly regulated genes in cardiomyocytes. There, only 23.3% of the downregulated genes were also decreased in mice. Most of them were either not regulated (48.2%) or even upregulated in mice (28.5%). The other cell types show a higher percentage of commonly downregulated genes.

We visualized all expression changes in one heatmap to further validate individual gene changes upon HFrEF (Fig. 4 A/B). Thereby, we found that around 30% of the genes show no changes in their expression upon heart failure (Fig. 4A, cluster 1). Most expression changes are consistently observable in all cell types (cluster 2-23) and therefore appear as general responses to injury which cannot be attributed to individual cell types. However, the remaining 16 clusters show cell type specific expression patterns (Fig. 4B). For example, cluster 25 holds a set of genes that show increased expression of genes in human FBs. Whereas cluster 28 in human ECs contains many genes that are down regulated. These changes are not detectable in other cell types for these genes and are therefore of utmost interest to follow up on specific gene expression changes in species specific cells. Similar patterns were found by observing commonly regulated genes (Fig. 4C). For humans, the largest number of DEGs were found in all cell types (1087 DEGs). The second largest groups contain DEGs that are found only in the individual cell types (Fig. 4C; Suppl. Fig. 4A). Thus, we identified 687 DEGs specific to human CM and 208 DEGs that can only be found in ECs. If we now determine the distribution of DEGs in mice, one finds larger populations of cell type specific genes and fewer, which are found in all populations (n = 228). Notably we found far fewer DEGs in the mouse SMCs than in the human samples. However, this could be related to the total number of SMCs in mice, which is far less in mice than in human samples (Fig. 1C & 3A). This could explain the lower number of DEGs found in all cell types. When we excluded SMCs from the common DEG population, we observed a similar number of DEGs in all cell types as in humans previously (Suppl. Fig. 4B).

Further, we analyzed the highest upregulated genes per cell type in humans and mice along with the regulation of that gene in the other species. Hereby, we observed how the genes with the largest changes in human heart failure patients behave in the respective mouse model (Suppl. Fig. 3).

We observed that the expression of the most regulated genes in human cell types show comparably less regulation in the mouse models. For example, we found *LDB2,* a gene of the LIM-Domain family, in human CMs as highly upregulated (Log2FC = 2.15) (Suppl. Fig. 3A). The LIM-Domain family genes are well known as adapter molecules that allow the assembly of transcriptional regulatory complexes in CM. However, in mice, *LDB2* is only mildly regulated upon HFrEF (Log2FC = 0.38). Other genes such as the VEGF receptor *FLT1,* which is upregulated in human cardiomyocytes, show a downregulation in mice CMs. This demonstrates that some genes have completely different expression patterns in humans and mice. However, some genes share similar regulation in their respective cell types. Thus, we observed that Phosphodiesterase 4D (*PDE4D)* and ADP Ribosylation Factor Like GTPase 15 (*ARL15)* show similar changes in ECs. Among the ten most upregulated genes in the mouse model data, we found three genes that also show a significant increase in their expression in humans (*RBPJ*, *SLC9A9*, *RUNX1*) (Suppl. Fig. 3B). The other genes, however, show little to no change. In contrast, if we investigate the expression changes in ECs, DEGs show an opposite direction in their expression change (*RBPJ*, *PID1*, *SLC9A9*). These differential gene expressions in the cell types suggest that some cell type specific responses may be different between human patients and mouse models.

Pathway enrichment results in cardiomyocytes

To address whether the relatively high number of significantly regulated genes indicate overall changes in pathways and pathological processes or whether the differences relate more to the alternative use of genes with similar functions in mice and humans, we further determined the implications for overall pathways in the individual cell types. Figure 4 shows a *simplifyEnrichment* heatmap cluster with word clouds of gene ontology terms regulated in

human or mouse cardiomyocytes. We generally observe more significantly enriched GSEA terms in humans than in mice (Fig. 5A). Important pathway terms regarding mitochondrial energy production and the electron chain were enriched in both species. Other terms involving developmental processes are enriched in humans compared to mice. Additionally, we investigated the set of genes found in cluster 25 and cluster 28 in more detail (Fig. 4B; Fig. 5B-C). Gene Ontology (GO) analysis on subsections of genes found in these clusters revealed a change in pathways associated with cell adhesion and extracellular processes (Fig. 5B). The second subsection of cluster 28 was associated with terms regarding cell differentiation processes, like "epithelial cell differentiation" or "angiogenesis" (Fig. 5C).

We identified cell type specifically regulated pathways upon HFrEF. Therefore, we investigated how the enriched signaling pathways differ between humans and mice in cardiomyocytes. We observed larger differences for pathways that were specifically regulated in humans. Among the most regulated pathways, specifically detected in the human, we found the terms "actin filament organization" and "angiogenesis" (Fig. 6A). Genes associated with these pathways were explicitly upregulated in patients (Fig. 6B). These gene sets are not found among the regulated pathways in mice (Suppl. Table 4). Examples of angiogenesis-related genes, which are specifically induced in human heart failure but not in mouse models, including receptors such as the *VEGF*-receptor FLT1, or transcription factors like the mesenchyme homeobox protein 2 (MEOX2) (Fig. 6B). In addition, many GTPase regulatory genes were found specifically increased in humans, including *MCF2L* and *RASGRF2*, which are known to regulate *RAC1*, and *SPATA13*, which enables guanyl-nucleotide exchange factor activity [34,35]. In contrast, we observed that signaling pathways mainly dealing with energy metabolism are commonly regulated in patients with heart disease as well as in mouse models. The genes included in pathways, such as "ATP biosynthetic process", "mitochondrial ATP synthesis", "aerobic electron transport chain" and "cellular respiration", show significant downregulation compared to their corresponding control (Fig. 6C). These data suggest conservation of disturbed mitochondrial metabolism in both mice and humans upon heart failure.

On the other hand, pathways such as "Wnt signaling pathway", "actin-myosin filament sliding" and "regulation of cell morphogenesis" are upregulated specifically in the mouse HFrEF model (Fig. 6A). Genes associated with Wnt signaling include *LRP6*, a known inhibitor of cardiomyocyte proliferation [36], and the serine/threonine-protein kinase *MARK2*, which regulates the stability of microtubules through phosphorylation and inactivation of several microtubule-associated proteins [36].

Furthermore, we repeated the GSEA analysis with the identified ECs in the human and mouse model data to gain further insight into the different cell types (Suppl. Fig. 5). Here, we found human-specific regulated terms such as "cardiac contraction" and "regulation of axonogenesis" (Suppl. Fig. 5A) that we only find in ECs but not in the previously analyzed CMs. The genes in these sets show a distinct regulation only observed in human data (Suppl. Fig. 5B). When we examined the commonly regulated metabolic pathways. We found similar terms and changes in gene expression related to impaired mitochondrial metabolism in EC as we had previously observed in CMs (Suppl. Fig. 5C). In ECs, we also found similar mouse-specific terms such as "cell morphogenesis" and the "Wnt signaling pathway", but also newly discovered pathways such as "positive regulation of steroid hormone secretion". Steroid hormones have been shown to coordinate microvascular function in obese mice endothelium [37]. Based on these results, one might speculate that this regulatory function is mouse specific. GSEA analysis for all other cell types can be found under [38]. All Source code for this study can be found in the paper specific Github repository [39].

Discussion

The ever growing number of published single cell experiments enables scientists to deepen the knowledge about transcriptional changes of individual cell types and species specific regulatory changes, upon disease conditions. Particular combination of single cell datasets from different species in the same UMAP projection allows the detection of well conserved or species specific regulatory networks [40–42].

Therefore, integrating datasets from different species with a well curated list of orthologues, has significant advantages and simplifies comparisons among species.

Here we propose *OrthoIntegrate,* an R-package that enables scientists to integrate single cell datasets from different species into a shared dimensional space. To generate high quality and uniquely mapped orthologous lists between different species, we implemented a new pipeline that increases the ono-to-one assignment of ontologies to improve single cell integration. Compared to the Ensembl orthologous list (Biomart), our pipeline results in up to 10% more uniquely assigned orthologues between human and mouse. Compared to the other databases OMA and InParanoid, OrthoIntegrate contained 8.6% and 9.3% more one-to-one orthologues (Fig. 3F).

OrthoIntegrate additionally contains functions that use the extended orthologous assignments to streamline the integration of single cell datasets from humans and mice. Moreover, it is highly adaptable and can be easily customized to support other species.

We demonstrated the usability of combining cross-species single cell data by using data sets of human and mouse heart failure with reduced ejection fraction.

In order to evaluate the species mixing and the biological conservation of different integration methods, we applied certain metrics from the scib package [32,43], which were also suggested by Song et.al [32,33,43]. The results are summarized in Figure 2D. We found that most batch correction scores improve by using OrthoIntegrate.

For biological conservation scores, we demonstrate that some metrics, like the "cell cycle conservation" are improved by using OrthoIntegrate. Which means that the variance caused by different cell cycle states of the cells is conserved via OrthoIntegrate. Other parameters like the NMI-score are reduced. But this score for example is strongly influenced by the cell type labeling [33], which was focused only on main cell type groups in these datasets, regardless of the existence of subpopulation or mixed cell type population clusters. In other words, subclusters of different cell types were not annotated in detail. Due to the increased numbers of features that are included in OrthoIntegrate, the clustering might be more diverged, likely

by species specific non-coding RNAs or other features, which are not included in the other databases. Therefore, the more divergent clustering, due the increased number of features in OrthoIntegrate combined with the broad cell type labeling might explain the slightly reduced NMI scores. However, since various publications have shown that long-non-coding RNAs have important regulatory roles in the heart [44–46], we think that these additional non-coding RNAs are an important resource to study species specific responses to different disease conditions, especially in the field of heart failure.

Commonly regulated pathways upon heart failure reflect an evolutionary conserved transcriptomic answer to severe damage in heart cells. One example is the conserved downregulation of critical mitochondrial metabolic pathways, which provide ATP for the heart (Fig. 5, Fig. 6A/C). As the heart is the most energy consuming organ, maintaining mitochondrial function plays a critical role and the decline in energy production limits heart function [47]. We could show that genes important for ATP biosynthesis and electron transport (e.g. *PGAM2*, *NDUFA1* and *TMEM126A*) are consistently downregulated in heart failure. *PGAM2* and *NDUFA1* have been described in the context of heart disease in mice [48] and rats [49], respectively, but their role in humans is unknown.

Besides commonly regulated pathways, we found differences between humans and mice upon heart failure. In cardiomyocytes, genes associated with "angiogenesis" were specifically enriched in humans. For example, the *VEGF* receptor *FLT1* was specifically increased in the human samples. *FLT1* primarily mediates *VEGF* signaling in endothelial cells, but its role in cardiomyocytes, besides high expression [50], is less clear [51]. Functionally, *FLT1* was shown to partially mediate *VEGF* induced cardiomyocyte differentiation [52] and regulate cardiomyocyte contractility in the embryonic zebrafish heart [53]. Cardiomyocyte specific deletion of *FLT1* was shown to worsen cardiac remodeling and hypertrophy induced by pressure overload [54], suggesting that the specific upregulation in humans may represent a compensatory cardioprotective mechanism that might not be conserved in mice.

A second human CM specific gene is *MEOX2*, which was assigned to "angiogenesis" because of its role in endothelial fatty acid transport [55]. *MEOX2* plays a critical role in development of all muscle lineages [56]. In cardiomyocytes, *MEOX2* overexpression blocks proliferation during heart morphogenesis [57]. All of these human CM specific genes have not been studied in mouse cardiomyocytes and their human specific regulation upon heart failure might be of utmost interest for future studies.

Among the pathways specifically enriched in mice we found predominant expression of genes associated with Wnt signaling. Although most identified genes have not been directly linked to cardiomyocyte-specific functions, Wnt signaling critically regulates cardiac hypertrophy, remodeling and regeneration [36,58]. Therefore, these findings and the other identified species specific pathways deserve more in depth validation and investigation.

To further demonstrate the functionality of OrthoIntegrate, we integrated scRNA-SEQ data from human [43], mouse [59] and zebrafish [59,60] brain tissue under Alzheimer's condition. Besides the evolutionary distance between these species, we could jointly cluster different cell types via OrthoIntegrate (Suppl. Fig 6 A-C) and detect commonly expressed marker genes within these cell clusters (Suppl. Fig 6 D-F).

In summary, our publicly available bioinformatic tool *OrthoIntegrate* simplifies the comparison of scRNA-SEQ datasets from humans and mice. Thereby we could identify conserved regulatory pathways upon heart failure. Furthermore, we identified cell type specific differences in both species. Also, we showed pathways like angiogenesis regulated explicitly in humans, and Wnt signaling pathways, specifically regulated in mice.

We anticipate that this study shows the benefits of the joint analysis of scRNA-SEQ data through *OrthoIntegrate.* Due to the growing number of scRNA-SEQ datasets, we hope that *OrthoIntegrate* encourages other scientists to perform comparative analysis between different species and thereby increasing knowledge about conserved or species specific pathway

responses in various diseases. This could improve the effective development of novel treatment strategies for heart failure or other diseases.

Limitations

The main limitation of our orthologue assignment and sample integration pipeline is the dependence on reliable databases for orthologous lists. Another problem with this approach is that it fails to consider the biological functions of the possible orthologues but selects the orthologue with highest sequence similarity. Second, our biological example has some limitations. While a decent number of healthy controls is available, the number of patients with HFrEF is limited. Knowing the biological heterogeneity of heart failure and comorbidities, variations are expected and the samples may not represent the representative and most common spectrum of heart failure. Finally, although the mouse model used is commonly applied in cardiovascular research, there are significant limitations due to the lack of underlying coronary artery disease and therapeutic pharmacological and interventions as it is done in humans. The integration of increasingly available published data both from alternative mice models and data derived from human samples will allow a refined comparative analysis in the future.

Methods

Study samples

The human heart samples used as controls were provided from the PRJEB39602 (Human Cell Atlas) project published in 2020. The heart tissue was obtained from deceased transplant organ donors who were between 45 and 70 years old and showed unremarkable cardiovascular history. The healthy mice samples (CTRL: n4-n9) were gathered by Vidal et al. (2019) and can be found using the Array Express Data Portal under E-MTAB-7869 (Supplement Table 1).

Heart samples from patients with heart failure with reduced ejection fraction (HFrEF) were gathered for this study from the Frankfurt University Hospital and subsequently processed at the Institute of Cardiovascular Regeneration (Frankfurt am Main, Germany), where the processed mice samples (CTRL: n1-n3, HFrEF: n1-n4) were also gathered and sourced. Nuclear isolation steps and single-nucleus RNA-sequencing library preparation were conducted as described in Nicin et al.; NCVR 2022 [61].

The human heart failure samples as well as the mice control and heart failure samples are published in Array Express with the accession E-MTAB-13264 (Supplement Table 1).

In order to provide another species and disease condition, we applied OrthoIntegrate on humans, mice and zebrafish in Alzheimer disease (AD) condition. Therefore we gathered scRNA-SEQ data from the prefrontal cortex (location matched) of human and mouse and zebrafish via scRead (human and mouse data, disease n=2; healthy n=2) [62] and GEO (GSE118577; n=3). The human and mouse samples originate from GSE129308 [43] and GSE143758 (AD) & GSE143758 (Healthy), respectively.

Single cell pre-processing

Single-cell RNA-seq results were processed by CellRanger (10x Genomics) version 6.1.1 software. The first step consisted of demultiplexing and processing raw base count files by the implemented *mkfastq* tool. The human raw reads were mapped to the reference genome hg38 (GRCh38-2020) using Cellranger count, whereas the mouse raw reads were mapped to the reference genome mm10 (GRCm38-2020). The secondary data analysis was conducted using the Seurat 4.1.0 package in R. The data sets were first combined into a Seurat object and then subjected to a filtering process. Barcodes with too low (< 300) or too high number of genes (> 6000) were sorted out and not considered further in the data analysis. In addition, barcodes with too low (< 500) and too high read counts (> 15000) were also sorted out. To further ensure no apoptotic cells or doublets were analyzed, we discarded barcodes with a

high percentage of mitochondrial content (> 5%). The filtered gene counts were then logarithmized and normalized according to the tutorial for data analysis with Seurat. Baseline characteristics for the samples can be found in Supplement Table 1.

Orthologue assignment and sample integration

In order to ensure the integration of single cell datasets from different species, we coded a function to assign animal model orthologues to the human nomenclature (or vice versa) using gene transfer format (GTF) files provided by Ensembl (GRCh38, GRCm38). In order to detect only well annotated genes between the species, predicted genes were removed. Afterwards orthologues to the human genes were determined using the R package biomaRt. This assigned the majority of genes in our human GTF file to at least one orthologue. If there were several entries of possible orthologues in the Ensembl database, a protein sequence comparison was initiated. Therefore, protein sequences were retrieved from the Uniprot database for the human gene and all possible orthologues in the second species. These sequences were then aligned using the R package Biostrings 2.60.2. The alignment score was calculated based on the Needleman-Wunsch global alignment algorithm [28] with substitution matrices. For nucleotide sequences, the *nucleotideSubstitutionMatrix* function was used to produce a substitution matrix for all IUPAC nucleic acid codes based upon match and mismatch parameters. BLOSUM50 matrix was retrieved from the NCBI Matrix Compendium for the protein sequence. The gene IDs with the highest amino acid sequence similarity between their canonical sequences were assigned. If there are no entries for canonical sequences in Uniprot, the nucleotide sequence similarity comparison is initiated. For this step, the unpredicted mRNA sequences for the gene in the first species and for the possible orthologues in the second species were obtained from the NCBI database and aligned analogously to the previous step. If no unpredicted mRNA is available for an entry, the function retrieves the unpredicted non coding RNA of the gene. This ensures that non coding genes without mRNAs can still be assigned correctly. In case both RNA sequences are not retrievable, predicted versions of mRNA and non coding RNA are retrieved. If all these

assignment steps are not successful, the Levenshtein distance was used to compare the ID symbols for possible orthologues and the orthologue with the lowest Levenshtein distance was selected.

Many long-non-coding RNAs are not listed in orthologue databases, therefore a final lowercase matching step was performed to assign genes like *Malat1* to the human *MALAT1*. With this globally applicable list of orthologues between species, the datasets were now filtered by these and then merged into one object using Seurat's canonical correlation analysis (CCA) integration.

Clustering, metrics calculation and annotation

To classify cells into clusters based on their expressed genes, we used the *FindNeighbors* and *FindClusters* (resolution parameter = 0.3) function implemented in Seurat. These clusters are determined by applying the shared nearest neighbors (SNN) clustering algorithm and the Uniform Manifold Approximation and Projection (UMAP) dimension reduction.

Calculations of the silhouette coefficient are based on computing a distance matrix based on the cell embeddings matrix for principal component analysis (PCA) performed by Seurat. This distance matrix includes the information of cell-cell distance, which is necessary for calculating the silhouette coefficient with our calculated clusters in the function *silhouette* of the cluster package (version 2.1.4). Additionally, the coefficients of the samples were averaged for each object. For applying the python scib package we converted our Seurat objects into Anndata objects using the zellkonverter package (version 1.10.1). We computed graph connectivity, principal component regression comparison, silhouette batch, kBET, LISI and cell cycle conservation scores for defining the species mixing score. Furthermore, the bio conservation score was calculated by computing the species type LISI, isolated labels F1 score, as well as the previously mentioned silhouette coefficient. The total score was then calculated by a weighted addition of species mixing score and bio conservation score (0.5 * species mixing score + 0.5 * bio conservation score). We provided the UniprotIDs of the orthologous lists obtained with the tools to be compared to the Orthology Benchmark web service to calculate

the Schlicker similarity scores for enzyme classification conservation and gene ontology conservation.

The orthologous lists for OMA, Biomart and InParanoid were created by following their introductions on their tool descriptions and by using the same GTF files as before (GRCh38, GRCm38).

For the assignment of cell clusters to cell types, we used a reference object that we had previously manually annotated with marker genes from Tombor et al. 2021 [63]. Here, the R package SingleR can be used to adopt marker genes that were used for the previous annotation of clusters of the reference object. These are then transferred and compared to marker genes of the cell clusters of our object to be annotated. Thus, a reproducible annotation can be guaranteed with the help of an exactly annotated data set.

Differential gene expression analysis and gene ontology analysis

Detection of differentially expressed genes (DEGs) for the cell type specific clusters was performed by the hurdle model of the MAST package (version 1.20.0). Results were filtered by their Bonferroni-adjusted p-value (p.adj < 0.05). The totality of DEGs were represented by Sankey plots created with the R package networkD3 (version 0.4). Additionally, bar plots were created using R package ggplot2, representing human DEGs and their regulation in mice. DEGs were also divided according to their species and cell type assignment and then visualized for DEGs with a positive Log2FC and separately in another plot, for DEGs with a negative Log2FC. Here, DEGs occurring in both human and mouse for the respective cell type have been pooled. Visualization was done in the form of a Circos plot (R package circlize 0.4.14). The gene regulation heatmap was created using the log2FC of all identified genes and a k-means clustering (k = 40) (R package ComplexHeatmap 2.16.0). Visualization of distinct and similar populations of genes in the analyzed cell types per species was achieved by creating venn diagrams with the Jvenn webtool.

Gene Set Enrichment Analysis (GSEA) was performed using the R package clusterProfiler (version 4.2.2) and the GO Database. GSEA terms were calculated separately for each cell type. The terms were sorted according to the Benjamini-Hochberg adjusted p.value and evaluated according to their "normalized enrichment distribution", which gives information about the regulation of the genes in the described pathway. A heatmap was created by clustering the GSEA terms by their similar geneIDs. (R package simplifyEnrichment 1.10.0). Additionally, the GSEA results were plotted in dot plots. Specifically, for genes described in the pathway, the standard error of the mean (SEM) bar plot was created (for their averaged UMIs) by using the R package ggplot2. Gene Ontology (GO) analyses were performed using the subsection of genes found in cluster 25 and cluster 28 as input for the webtool Metascape.

Funding

List of abbreviations

RNA: ribonucleic acid; mRNA: messenger RNA; scRNA-SEQ: single cell RNA sequencing; HFrEF: heart failure patients with reduced ejection fraction; LAD: coronary artery model; CM: cardiomyocyte; PC: pericytes; SMC: smooth muscle cell; FB: fibroblast; EC: endothelial cell; IC: immune cell; NC: neuronal cell; DEG: differentially expressed gene; GO: gene ontology; AD: Alzheimer disease; GTF: gene transfer format; CCA: canonical correlation analysis; SNN: shared nearest neighbors; UMAP: uniform manifold approximation and projection; PCA: principal component analysis; GSEA: gene set enrichment analysis; SEM: standard error of the mean; GO: gene ontology; HLC: human cell atlas; GC: graph connectivity; PCR: principal component regression comparison; bASW: batch average silhouette width; CCC: cell cycle

conservation; NMI: normalized mutual information; ILF1: isolated labels F1 score; SC: silhouette coefficient; SMS: species mixing score; BCS: bio conservation score; NES: normalized enrichment score; UMI: unique molecular identifier.

Author Contributions

M.R.J. and D.J. performed bioinformatic analyses, and data analysis. L.T. and M.H.S. supported bioinformatic analysis. M.A., T.H., F.E. T.W., A.M.Z. provided human samples and patient data. A.F. performed mouse breeding and sample preparation. W.A. conducted single-cell library generation. D.J and S.D. designed experiments. M.R.J., S.D. and D.J. wrote the manuscript.

Availability of source code and requirements

Project name: OrthoIntegrate

Project home page: https://github.com/MarianoRuzJurado/OrthoIntegrate [38]

Operating system(s): Platform independent

Programming language: R

Other requirements: certain R-packages (Seurat (>= 4.2.0), ggplot2 (>= 3.3.6), ggpubr (>= 0.4.0), biomaRt (>= 2.52.0), rtracklayer (>= 1.56.1), mygene (>= 1.32.0), UniprotR (>= 2.2.2), RecordLinkage (>= 0.4-12.3), Biostrings (>= 2.64.1), rentrez (>= 1.2.3), stringr (>= 1.4.1), svglite (>= 2.1.0), dplyr (>= 1.1.2), tidyr (>= 1.3.0)

License: GNU GPL

Any restrictions to use by non-academics: none

RRID: **SCR_025029, OrthoIntegrate**

Data availability

The single nuclei data for humans have been deposited in the Human Cell Atlas (HLC) database and can be accessed through the HCA Data Portal [64]. The mice sequencing data are available through ArrayExpress under the accession number E-MTAB-7869. All supporting data and materials are available in the *GigaScience* GigaDB database [65].

Ethics declaration

The authors declare no competing interests.

References

1. Ericsson AC, Crim MJ, Franklin CL. A brief history of animal modeling. *Mo Med.* 2013; 110(3): 201-205. 2. Wall RJ, Shani M. Are animal models as good as we think? *Theriogenology*. 2008; doi: 10.1016/j.theriogenology.2007.09.030.

3. Shanks N, Greek R, Greek J. Are animal models predictive for humans? *Philos Ethics Humanit Med*. 2009; doi: 10.1186/1747-5341-4-2.

4. Uhl EW, Warner NJ. Mouse Models as Predictors of Human Responses: Evolutionary Medicine. *Curr Pathobiol Rep*. 2015; doi: 10.1007/s40139-015-0086-y.

5. . The Mouse in Biomedical Research: History, Wild Mice, and Genetics. Elsevier; 2006.

6. Riehle C, Bauersachs J. Small animal models of heart failure. *Cardiovasc Res*. 2019; doi: 10.1093/cvr/cvz161.

7. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al.. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002; doi: 10.1038/nature01262.

8. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Mouse Genome Database Group. Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*. 2019; doi: 10.1093/nar/gky1056.

9. Noll NA, Lal H, Merryman WD. Mouse Models of Heart Failure with Preserved or Reduced Ejection Fraction. *Am J Pathol*. 2020; doi: 10.1016/j.ajpath.2020.04.006.

10. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. *Nat Rev Genet*. 2017; doi: 10.1038/nrg.2017.19.

11. Dayeh NR, Tardif J-C, Shi Y, Tanguay M, Ledoux J, Dupuis J. Echocardiographic validation of pulmonary hypertension due to heart failure with reduced ejection fraction in mice. *Sci Rep*. 2018; doi: 10.1038/s41598-018-19625-2.

12. Sawall S, Franke D, Kirchherr A, Beckendorf J, Kuntz J, Maier J, et al.. In Vivo Quantification of Myocardial Infarction in Mice Using Micro-CT and a Novel Blood Pool Agent. *Contrast Media Mol Imaging*. 2017; doi: 10.1155/2017/2617047.

13. Van Craeyveld E, Jacobs F, Gordts SC, De Geest B. Low-density lipoprotein receptor gene transfer in hypercholesterolemic mice improves cardiac function after myocardial infarction. *Gene Ther*. 2012; doi: 10.1038/gt.2011.147.

14. Swynghedauw B. Molecular mechanisms of myocardial remodeling. *Physiol Rev*. 1999; doi: 10.1152/physrev.1999.79.1.215.

15. Ertl G, Frantz S. Healing after myocardial infarction. *Cardiovasc Res*. 2005; doi: 10.1016/j.cardiores.2005.01.011.

16. Vigen R, Maddox TM, Allen LA. Aging of the United States population: impact on heart failure. *Curr Heart Fail Rep*. 2012; doi: 10.1007/s11897-012-0114-8.

17. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. 2022; doi: 10.1002/ctm2.694.

18. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al.. Comprehensive Integration of Single-Cell Data. *Cell*. 2019; doi: 10.1016/j.cell.2019.05.031.

19. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018; doi: 10.1038/s12276-018-0071-8.

20. Lu Y, Rosenfeld R, Nau GJ, Bar-Joseph Z. Cross species expression analysis of innate immune response. *J Comput Biol*. 2010; doi: 10.1089/cmb.2009.0147.

21. Kristiansson E, Österlund T, Gunnarsson L, Arne G, Larsson DGJ, Nerman O. A novel method for cross-species gene expression analysis. *BMC Bioinformatics*. 2013; doi: 10.1186/1471-2105-14-70.

22. Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, et al.. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A*. 2013; doi: 10.1073/pnas.1222878110.

23. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*. 2018; doi: 10.1093/gigascience/giy059.

24. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al.. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. Nature Publishing Group; 2019; doi: 10.1038/s41592-019-0619-0.

25. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al.. Ensembl 2022. *Nucleic Acids Res*. 2022; doi: 10.1093/nar/gkab1049.

26. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al.. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022; doi: 10.1093/nar/gkab1112.

27. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021; doi: 10.1093/nar/gkaa1100.

28. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970; doi: 10.1016/0022-

2836(70)90057-4.

29. Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, et al.. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* 2021; doi: 10.1093/nar/gkaa1007.

30. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al.. BioMart--biological queries made easy. *BMC Genomics.* 2009; doi: 10.1186/1471-2164-10-22.

31. O'Brien KP, Remm M, Sonnhammer ELL. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 2005; doi: 10.1093/nar/gki107.

32. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al.. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods.* 2022; doi: 10.1038/s41592-021-01336-8.

33. Song Y, Miao Z, Brazma A, Papatheodorou I. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nat Commun.* 2023; doi: 10.1038/s41467-023-41855-w.

34. Kawasaki Y, Sagara M, Shibata Y, Shirouzu M, Yokoyama S, Akiyama T. Identification and characterization of Asef2, a guanine-nucleotide exchange factor specific for Rac1 and Cdc42. *Oncogene.* 2007; doi: 10.1038/sj.onc.1210574.

35. Huang S-C, Chen Y-M, Hu Y-Y, Shi Y-J, Xiao Q-W, Li Z, et al.. Downregulation of MCF2L Promoted the Ferroptosis of Hepatocellular Carcinoma Cells through PI3K/mTOR Pathway in a RhoA/Rac1 Dependent Manner. *Dis Markers.* 2022; doi: 10.1155/2022/6138941.

36. Wu Y, Zhou L, Liu H, Duan R, Zhou H, Zhang F, et al.. LRP6 downregulation promotes cardiomyocyte proliferation and heart regeneration. *Cell Res.* 2021; doi: 10.1038/s41422-020-00411-7.

37. Biwer LA, Carvajal BV, Lu Q, Man JJ, Jaffe IZ. Mineralocorticoid and Estrogen Receptors in Endothelial Cells Coordinately Regulate Microvascular Function in Obese Female Mice. *Hypertension.* 2021; doi: 10.1161/HYPERTENSIONAHA.120.16911.

38. Jurado MR. MarianoRuzJurado/OrthoIntegrate: OrthoIntegrate_v1.0.1. Zenodo; 2024. https://doi.org/10.5281/zenodo.10686731

39. Jurado MR. MarianoRuzJurado/RuzJurado_et_al_2023: RuzJurado_et_al_2023. Zenodo; https://github.com/MarianoRuzJurado/RuzJurado_et_al_2023

40. Balachandran S, Pozojevic J, Sreenivasan VKA, Spielmann M. Comparative single-cell analysis of the adult heart and coronary vasculature. *Mamm Genome.* 2022; doi: 10.1007/s00335-022-09968-7.

41. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018; doi: 10.1038/nbt.4096.

42. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al.. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* 2016; doi: 10.1016/j.cels.2016.08.011.

43. Otero-Garcia M, Mahajani SU, Wakhloo D, Tang W, Xue YQ, Morabito S, Pan J, Oberhauser J, Madira AE, Shakouri T, Deng Y. Molecular signatures underlying

neurofibrillary tangle susceptibility in Alzheimer's disease. Neuron. 2022 Sep 21;110(18):2929-48. https://doi.org/10.1016/j.neuron.2022.06.021

44. Neumann P, Jaé N, Knau A, Glaser SF, Fouani Y, Rossbach O, et al.. The lncRNA GATA6-AS epigenetically regulates endothelial gene expression via interaction with LOXL2. *Nat Commun.* 2018; doi: 10.1038/s41467-017-02431-1.

45. Fouani Y, Kirchhof L, Stanicek L, Luxán G, Heumüller AW, Knau A, et al.. The splicing-regulatory lncRNA NTRAS sustains vascular integrity. *EMBO Rep.* 2022; doi: 10.15252/embr.202154157.

46. Michalik KM, You X, Manavski Y, Doddaballapur A, Zörnig M, Braun T, et al.. Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth. *Circ Res.* 2014; doi: 10.1161/CIRCRESAHA.114.303265.

47. Huss JM, Kelly DP. Mitochondrial energy metabolism in heart failure: a question of balance. *J Clin Invest.* 2005; doi: 10.1172/JCI24405.

48. Okuda J, Niizuma S, Shioi T, Kato T, Inuzuka Y, Kawashima T, et al.. Persistent overexpression of phosphoglycerate mutase, a glycolytic enzyme, modifies energy metabolism and reduces stress resistance of heart in mice. *PLoS One.* 2013; doi: 10.1371/journal.pone.0072173.

49. Liu T, Chen L, Kim E, Tran D, Phinney BS, Knowlton AA. Mitochondrial proteome remodeling in ischemic heart failure. *Life Sci.* 2014; doi: 10.1016/j.lfs.2014.02.004.

50. Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al.. A single-cell type transcriptomics map of human tissues. *Sci Adv.* 2021; doi: 10.1126/sciadv.abh2169.

51. Kurotsu S, Osakabe R, Isomi M, Tamura F, Sadahiro T, Muraoka N, et al.. Distinct expression patterns of Flk1 and Flt1 in the coronary vascular system during development and after myocardial infarction. *Biochem Biophys Res Commun.* 2018; doi: 10.1016/j.bbrc.2017.11.094.

52. Chen Y, Amende I, Hampton TG, Yang Y, Ke Q, Min J-Y, et al.. Vascular endothelial growth factor promotes cardiomyocyte differentiation of embryonic stem cells. *Am J Physiol Heart Circ Physiol.* 2006; doi: 10.1152/ajpheart.00363.2005.

53. Rottbauer W, Just S, Wessels G, Trano N, Most P, Katus HA, et al.. VEGF-PLCgamma1 pathway controls cardiac contractility in the embryonic heart. *Genes Dev.* 2005; doi: 10.1101/gad.1319405.

54. Mei L, Huang Y, Lin J, Chu M, Hu C, Zhou N, et al.. Increased cardiac remodeling in cardiac-specific Flt-1 receptor knockout mice with pressure overload. *Cell Tissue Res.* 2015; doi: 10.1007/s00441-015-2209-5.

55. Coppiello G, Collantes M, Sirerol-Piquer MS, Vandenwijngaert S, Schoors S, Swinnen M, et al.. Meox2/Tcf15 heterodimers program the heart capillary endothelium for cardiac fatty acid uptake. *Circulation.* 2015; doi: 10.1161/CIRCULATIONAHA.114.013721.

56. Skopicki HA, Lyons GE, Schatteman G, Smith RC, Andrés V, Schirm S, et al.. Embryonic expression of the Gax homeodomain protein in cardiac, smooth, and skeletal muscle. *Circ Res.* 1997; doi: 10.1161/01.res.80.4.452.

57. Fisher SA, Siwik E, Branellec D, Walsh K, Watanabe M. Forced expression of the homeodomain protein Gax inhibits cardiomyocyte proliferation and perturbs heart

morphogenesis. *Development*. 1997; doi: 10.1242/dev.124.21.4405.

58. Bergmann MW. WNT signaling in adult cardiac hypertrophy and remodeling: lessons learned from cardiac development. *Circ Res*. 2010; doi: 10.1161/CIRCRESAHA.110.223768.

59. Habib N, McCabe C, Medina S, Varshavsky M, Kitsberg D, Dvir-Szternfeld R, et al.. Disease-associated astrocytes in Alzheimer's disease and aging. *Nat Neurosci*. 2020; doi: 10.1038/s41593-020-0624-8.

60. Cosacak MI, Bhattarai P, Reinhardt S, Petzold A, Dahl A, Zhang Y, et al.. Single-Cell Transcriptomics Analyses of Neural Stem Cell Heterogeneity and Contextual Plasticity in a Zebrafish Brain Model of Amyloid Toxicity. *Cell Rep*. 2019; doi: 10.1016/j.celrep.2019.03.090.

61. Nicin L, Schroeter SM, Glaser SF, Schulze-Brüning R, Pham M-D, Hille SS, et al.. A human cell atlas of the pressure-induced hypertrophic heart. *Nature Cardiovascular Research*. Nature Publishing Group; 2022; doi: 10.1038/s44161-022-00019-7.

62. Jiang J, Wang C, Qi R, Fu H, Ma Q. scREAD: A Single-Cell RNA-Seq Database for Alzheimer's Disease. *iScience*. 2020; doi: 10.1016/j.isci.2020.101769.

63. Tombor LS, John D, Glaser SF, Luxán G, Forte E, Furtado M, et al.. Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction. *Nat Commun*. 2021; doi: 10.1038/s41467-021-20905-1.

64. Litviňuková M, Talavera-López C, Maatz H, Reichart D, Worth CL, Lindberg EL, et al.. Cells of the adult human heart. Nature. 2020; doi: 10.1038/s41586-020-2797-4; URL: https://www.heartcellatlas.org.

65. Jurado MR; Tombor LS; Arsalan M; Holubec T; Emrich F; Walther T; Abplanalp W; Fischer A; Zeiher AM; Schulz MH; Dimmeler S; John D. Supporting data for "Improved integration of single cell transcriptome data demonstrates common and unique signatures of heart failure in mice and humans" GigaScience Database 2024. http://dx.doi.org/10.5524/102508

Figure Legends

Fig.1: Integrated human and mouse snRNA-SEQ data of healthy and heart failure samples.

(A) Use case diagram of *OrthoIntegrate*: Shown are the steps that are run by the user within their standard Seurat workflow. First the Import function is used to create Seurat objects from scRNA-seq data, second orthologues are searched by the BuildOrtholog function and the third step creates an integrated object with uniform nomenclature by using the IntegrateObjects. (B) UMAP showing human cells (red) and mice cells (blue) in a common UMAP projection. In addition, cell types for the cell clusters can be seen. (C) UMAP with defined clusters according to Seurat's clustering, divided by species. Cells of mouse and human origin commingled in all clusters. There are no clusters formed that originated from only one of the two species. The cells were identified as cardiomyocytes (red), fibroblasts (yellow), endothelial cells (green), pericytes (turquoise), immune cells (blue), smooth muscle cells (purple) and neuronal cells (pink). (D) Bar plot showing cell composition of cell types in human (red) and mice (blue) samples. Samples were grouped based on their origin into human controls from the left ventricle (Human-CTRLlv), human HFrEF (Human-HFrEF), mouse controls (Mice-CTRL), and mouse HFrEF model (Mice-HFrEF). Cell types were then analyzed for their composition from the previously mentioned groups and plotted. P-values above the certain groups were calculated by two-sided Student's t-test. (E) Dot plot depicting the average expression levels and expression proportions in human samples of the top ten feature genes for the found cell types. The size of the dot represents the proportion of cells expressing the indicated gene within a cell type, and the color indicates the average expression level of cells. (F) Dot plot depicting the average expression levels and expression proportions in mice samples of the top ten feature genes for the found cell types. Similar to (E) the size of the dot represents the proportion of cells expressing the indicated gene within a cell type, and the color indicates the average expression level of cells.

Fig. 2: Comparison of snRA-Seq data integration with orthologues from *OrthoIntegrate* and other orthologue databases. (A) Box plot showing the average silhouette coefficient for clusterings based on different databases and tools. The dark blue box stands for the silhouette coefficient of the clustering made with an orthologous list using the tool OMA (Orthologous matrix). It is followed by the results for biomaRt (light blue), InParanoid (green) and the pipeline *OrthoIntegrate* (yellow). On the y-axis you can see the value of the silhouette coefficient. Additionally, each silhouette coefficient was calculated for each sample and depicted as a circle in their species specific color. (B) Bar plot with number of orthologues found which codes for a protein (C) and bar plot with number of orthologues found which codes for lncRNA. On the x-axis the used tool is depicted. (D) Table showing results of different metric calculations to comprehend batch correction and biological conservation of clusterings based on orthologous lists of OMA, biomaRt, InParanoid and *OrthoIntegrate* (GC = graph connectivity, PCR = principal component regression comparison, bASW = batch average silhouette width, CCC = cell cycle conservation, NMI = normalized mutual information, ILF1 = isolated labels F1 score, SC = silhouette coefficient, SMS = species mixing score, BCS = bio conservation score). The color code represents low and high values and is scaled per column (low = green, brown; high = blue, yellow). (E) Schlicker similarity scores calculated for OMA (red), Biomart (green), InParanoid (blue) and OrthoIntegrate (purple) in terms of enzyme classification conservation (left) and gene ontology conservation (right). (F) Venn Diagram highlighting the numbers of uniquely found orthologues between human and mouse per tool and their overlap between each other (blue = OMA, red = biomart, green = biomart, yellow = *OrthoIntegrate*).

Fig.3: Similarities and differences revealed by DEG analysis

(A) Sankey plot illustrating the distribution of differentially regulated genes (DEG) in the corresponding cell types. The width of the paths illustrates the number of DEGs that are either human specific (yellow), detected in both species (light green) or mouse-specific (dark green). DEG analysis was performed for each cell type individually. Neuronal cells were omitted from all further analyses due to their insufficient number of cells in the mouse data. (B) Bar graph

of up (top) and down (bottom) regulated genes in humans, along with the expression in mice. The panels show genes that are either commonly regulated (left), regulated in humans and not regulated in mice (middle) and regulated in opposite directions.

Fig.4: DEG analysis shows similar and different populations of regulation in gene expression patterns upon heart failure in humans and mice. (A) Heatmap of log2FC values (Control vs HFrEF) for all genes and all cell types. The y-axis describes all genes (16.545) clustered by a k-means algorithm (k = 40). The x-axis shows the species and the additional clustering into the different cell types. Positive log2FCs are represented by a red color, while negative scores give a blue color. (B) Close-up of the 24-40 k-means clusters of log2FCs of genes in which most cell type-specific differences are observed. (C) Venn Diagrams of all identified DEGs in human (top) and mouse (bottom) (log2FC > 0.1 and p-adjusted < 0.05).

Fig.5: GSEA analysis reveals more regulated pathways in heart failure in human cardiomyocytes than in mice, with the terms found sharing many keywords. (A) Heatmap clustering significant GSEA results (p.adj < 0.25) of DEGs found in human and mouse cardiomyocytes by similar GeneIDs in the pathways. Bar graphs are shown on the left y-axis representing the number of pathways found in the respective cluster for the given species and condition. In addition, the adjusted p-value is color-coded from 1 (green) to the smallest p-value found ~0.025 (red). On the right side of the y-axis keywords describing the found pathways in that cluster are shown, where the size of the word represents its frequency in the terms (larger = most, smaller = less). (B) Bar graph showing the first 20 GO terms found by analyzing genes in cluster 25. Terms were sorted by their logarithmized and Bonferroni-adjusted p-values resulting in high significant pathways depicted first (p.adj < 0.05). (C) Bar graph similar to (B) with terms found in a subsection of genes in cluster 28.

Fig.6: Common and distinct regulated pathways found in human and mouse cardiomyocytes.

(A) Dot plot visualizing the ten most significant pathways for terms only to be found regulated in humans, commonly regulated and specific in mice. The size of the dots corresponds to the negative log10 of the Benjamini Hochberg adjusted p-value and the color-code represents the normalized enrichment score (NES), with upregulated pathways shown in red and downregulated pathways in blue. The y-axis depicts the description of the identified term.

(B) Bar plot with mean values for the amount of unique molecular identifiers (UMIs) in the cells for the shown genes. The genes are identified to be dissimilarly regulated between humans and mice for pathways specifically found in humans. (C) Bar graph similar to (B) with mean values for UMIs in cells for genes downregulated in both species for commonly found terms.

(D) Bar graph similar to (B) and (C) with mean values for UMIs in cells for genes which are uniquely found to be regulated in terms specifically identified in mice. P-values above the certain groups were calculated by two-sided Student's t-test.

Supplement Fig.1: Integration process of human/mouse snRNA-SEQ data.

(A) UML-Activity-Flowchart showing orthologue assignment pipeline for human to mouse gene symbols. First, the Gene transfer format file (GTF) for humans (GRCh38) is used to get all annotated gene nomenclatures. Then all genes are filtered out which are only predicted and not clearly detected. This list is now searched for orthologues using the Ensembl database; all 1:1 assignments can be included in our orthologous list. In the case of multiple assignments, all possible replacements are examined according to their protein sequence and an alignment score is calculated according to the global sequence alignment. If there is no protein sequence in the Uniprot database, the alignment score is calculated based on the nucleotide sequence using the NCBI database. Now the gene with the best result is set as an orthologue. All unassigned genes are additionally compared with the GTF file of GRCm38 using a lowercase matching and if there is a match, they will be added to the orthologue list. If all these approaches for a gene do not result in an orthologue, a Levenshtein distance score is

calculated based on their gene names. (B) Single cell integration pipeline showing steps performed to integrate human and mouse scRNA-SEQ data in a joined UMAP projection. The scRNA-SEQ data from our human and mouse samples are first converted into Seurat objects and normalized. After that, clustering takes place and cell types can be determined. Using the orthologous list from our orthologue assignment algorithm, the objects can be subsetted according to the genes found and their nomenclature unified. This is followed by an integration into a single object and a clustering step.

Supplement Fig. 2: Overlapping of human and mouse cells after Seurat integration with tool specific orthologous list. (A) UMAPs showing human cells (red) and mice cells (blue) in a common UMAP projection for each tool used for integrating the data. First UMAP was performed on an object made with an orthologous list of OMA, followed by Biomart and InParanoid. The last UMAP shows the projection for the *OrthoIntegrate* pipeline. (B) Similar to (A), UMAPs are shown to visualize the clustering created with the R package Seurat using the Louvain algorithm. The objects are found in the same order as previously described. (C) Similar to (A) and (B), UMAPs are shown with the cell type annotation. The order is maintained as in (A) and (B) (same parameters are used for each UMAP; resolution = 0.3).

Supplement Fig. 3: Circos plots of DEGs show specific and similar expressed DEGs.
(A) Circos plot showing the ten most upregulated genes in human HFrEF (log2FC), separated for all cell types. Red line indicates the log2FC for human DEGs, while the blue line indicates the log2FC of the corresponding mouse gene. (B) Circos plot similar to (A) illustrates the ten most upregulated genes in mice HFrEF samples in comparison to the regulation of these genes in humans.

Supplement Fig. 4: DEG analysis shows population of shared and cell type-specific genes across cell types in humans and mice.

(A) Upset plot of Human DEGs found in the analyzed cell types. The groups are sorted by their intersection size and plotted on the x-axis of the plot. Additionally the overlapping groups are visualized by a connected dot plot. The size of the DEGs per cell type are shown on the y-axis.

(B) Upset plot of Mouse DEG populations similar to previous plot in (A).

Supplement Fig. 5: GSEA analysis shows regulated pathways upon heart failure in human and mouse endothelial cells.

(A) Dot plot visualizing the ten most significant pathways for terms only to be found regulated in humans, commonly regulated and specific in mice endothelial cells. The size of the dots corresponds to the negative log10 of the Benjamini Hochberg adjusted p-value and the color-code represents the normalized enrichment score (NES), with upregulated pathways shown in red and downregulated pathways in blue. The y-axis depicts the description of the identified term.

(B) Bar plot with mean values for the amount of unique molecular identifiers (UMIs) in the cells for the shown genes. The genes are identified to be dissimilarly regulated between humans and mice for pathways specifically found in humans. (C) Bar graph similar to (B) with mean values for UMIs in cells for genes downregulated in both species for commonly found terms.

(D) Bar graph similar to (B) and (C) with mean values for UMIs in cells for genes which are uniquely found to be regulated in terms specifically identified in mice. P-values above the certain groups were calculated by two-sided Student's t-test.


Supplement Fig. 6: Integration of human, mouse and zebrafish scRNA-SEQ prefrontal cortex data of healthy and Alzheimer patients.

(A) UMAP showing human cells (red), mice cells (blue) and zebrafish cells (green) in a common UMAP projection. (B) UMAP with defined clusters according to Seurat's clustering.

Cells of mouse, human and zebrafish origin commingled in the astrocyte cluster (green). Excitatory (red) and inhibitory (green) neurons are mostly present in human data. (C) Bar plot showing cell composition of cell types in human (red), mice (blue) and zebrafish (green) samples. Samples were grouped based on their origin into human, mouse and zebrafish controls from the prefrontal cortex (Healthy) and Alzheimer disease human, mouse and zebrafish (Alzheimer). Cell types were then analyzed for their composition from the previously mentioned groups and plotted. (D) Dot plot depicting the average expression levels and expression proportions in human samples of the top fifteen feature genes for the found cell types. The size of the dot represents the proportion of cells expressing the indicated gene within a cell type, and the color indicates the avera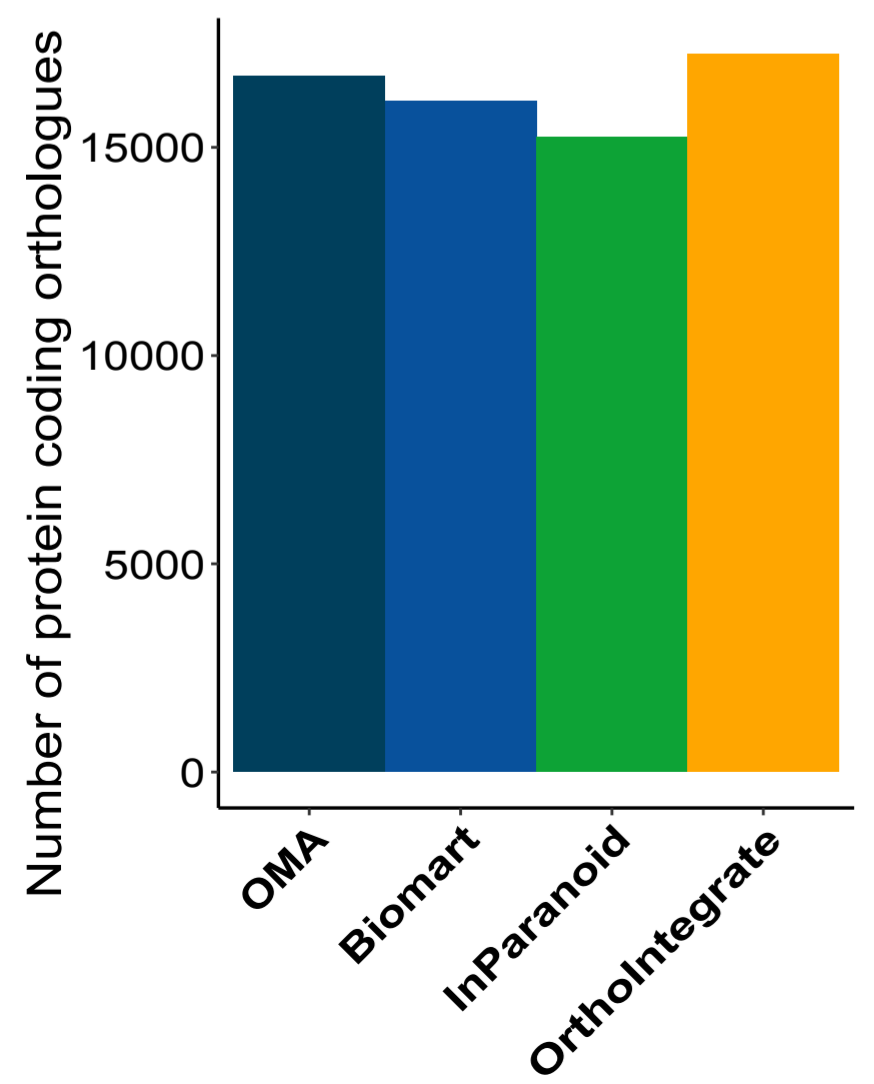ge expression level of cells. (E) Dot plot depicting the average expression levels and expression proportions in mice samples of the top fifteen feature genes for the found cell types. Similar to (D) the size of the dot represents the proportion of cells expressing the indicated gene within a cell type, and the color indicates the average expression level of cells. (F) Dot plot depicting the average expression levels and expression proportions in zebrafish samples of the top fifteen feature genes for the found cell types. Similar to (D and E) the size of the dot represents the proportion of cells expressing the indicated gene within a cell type, and the color indicates the average expression level of cells.

Fig_1

Click here to access/download;Figure;Fig_1.pdf ⬙
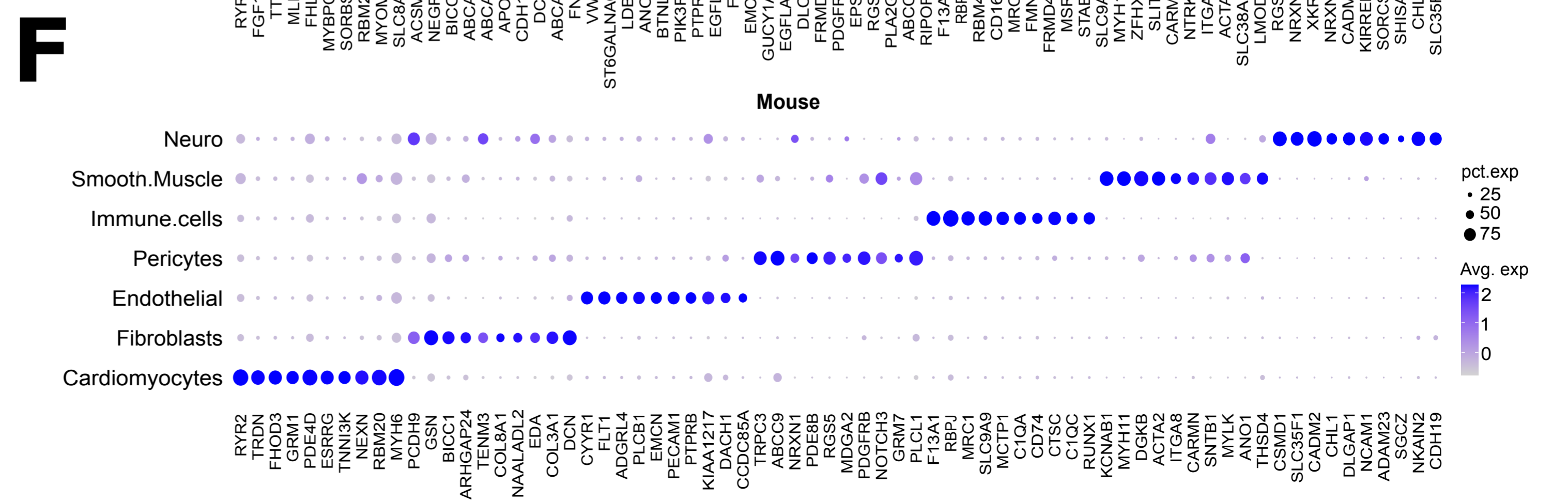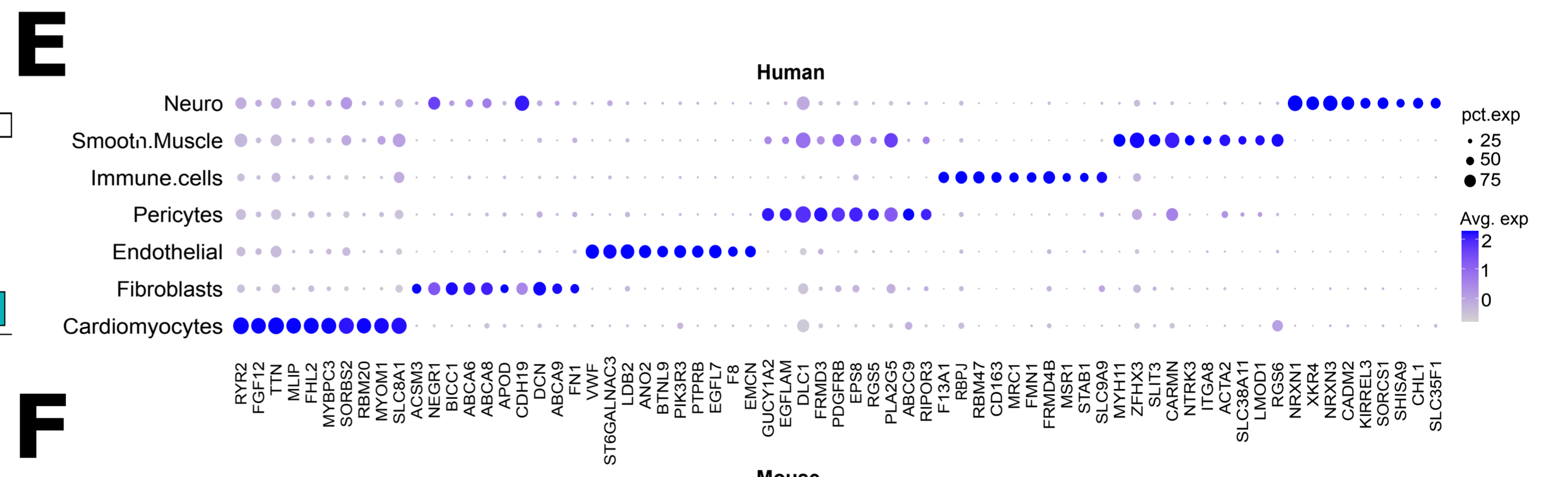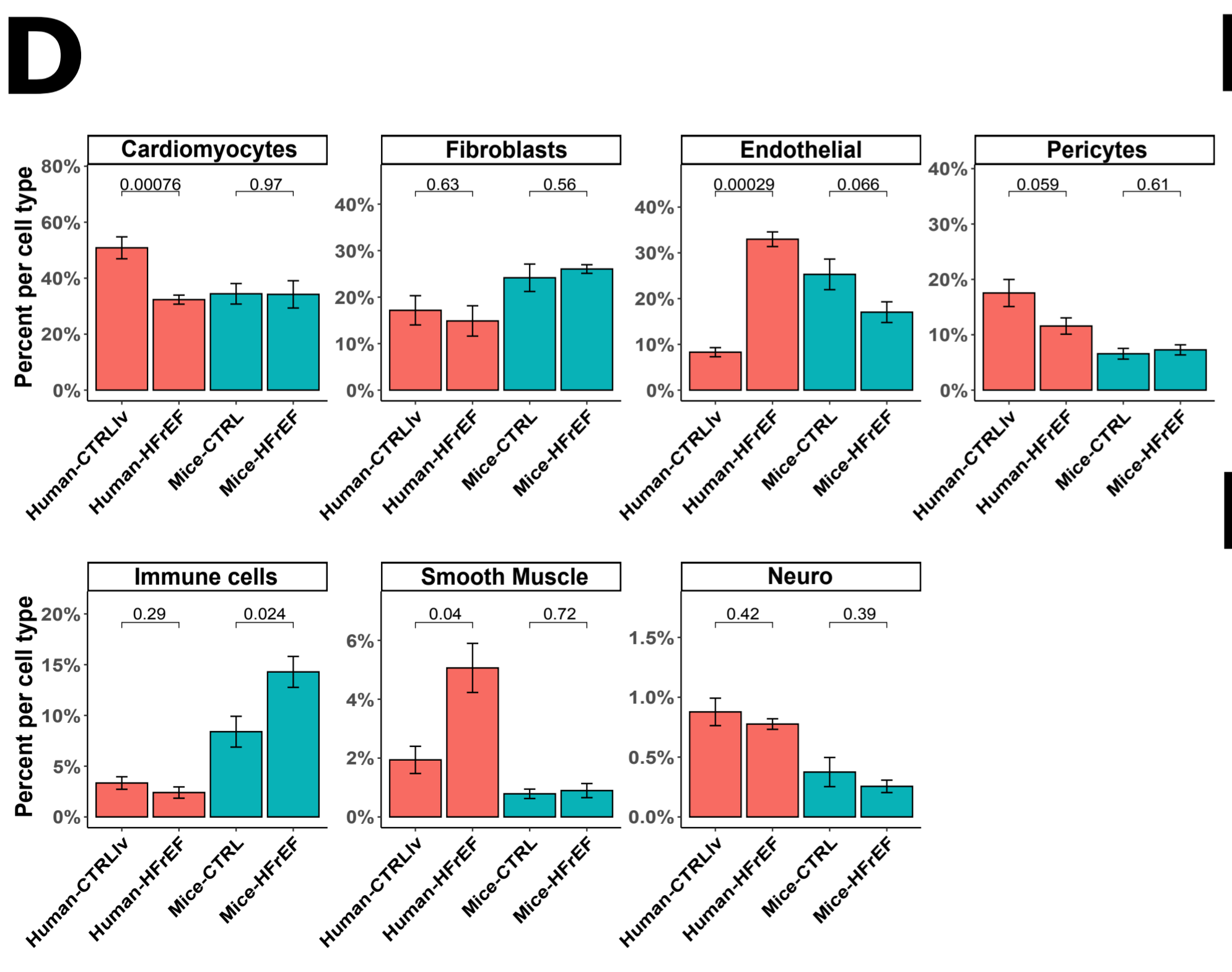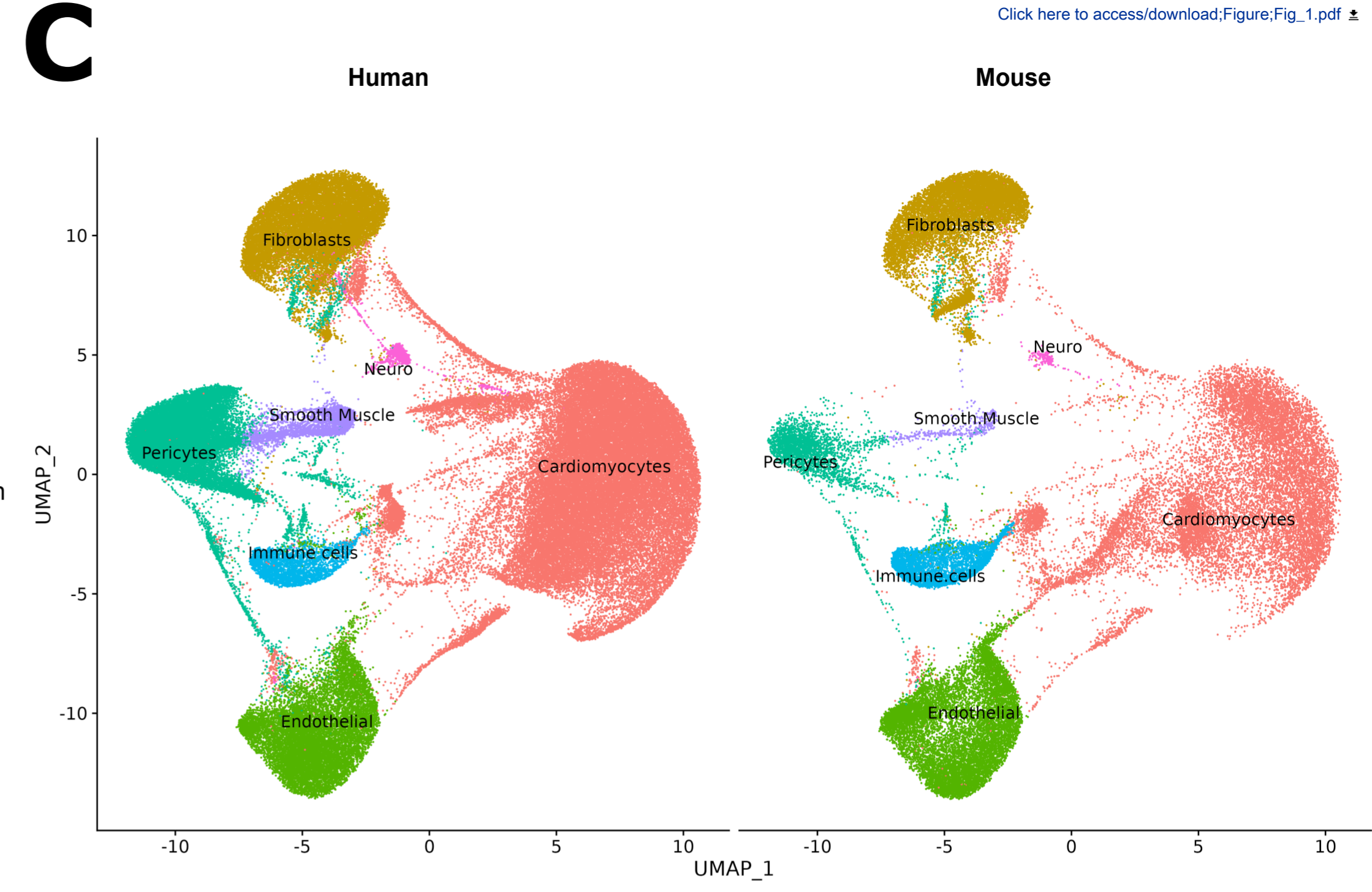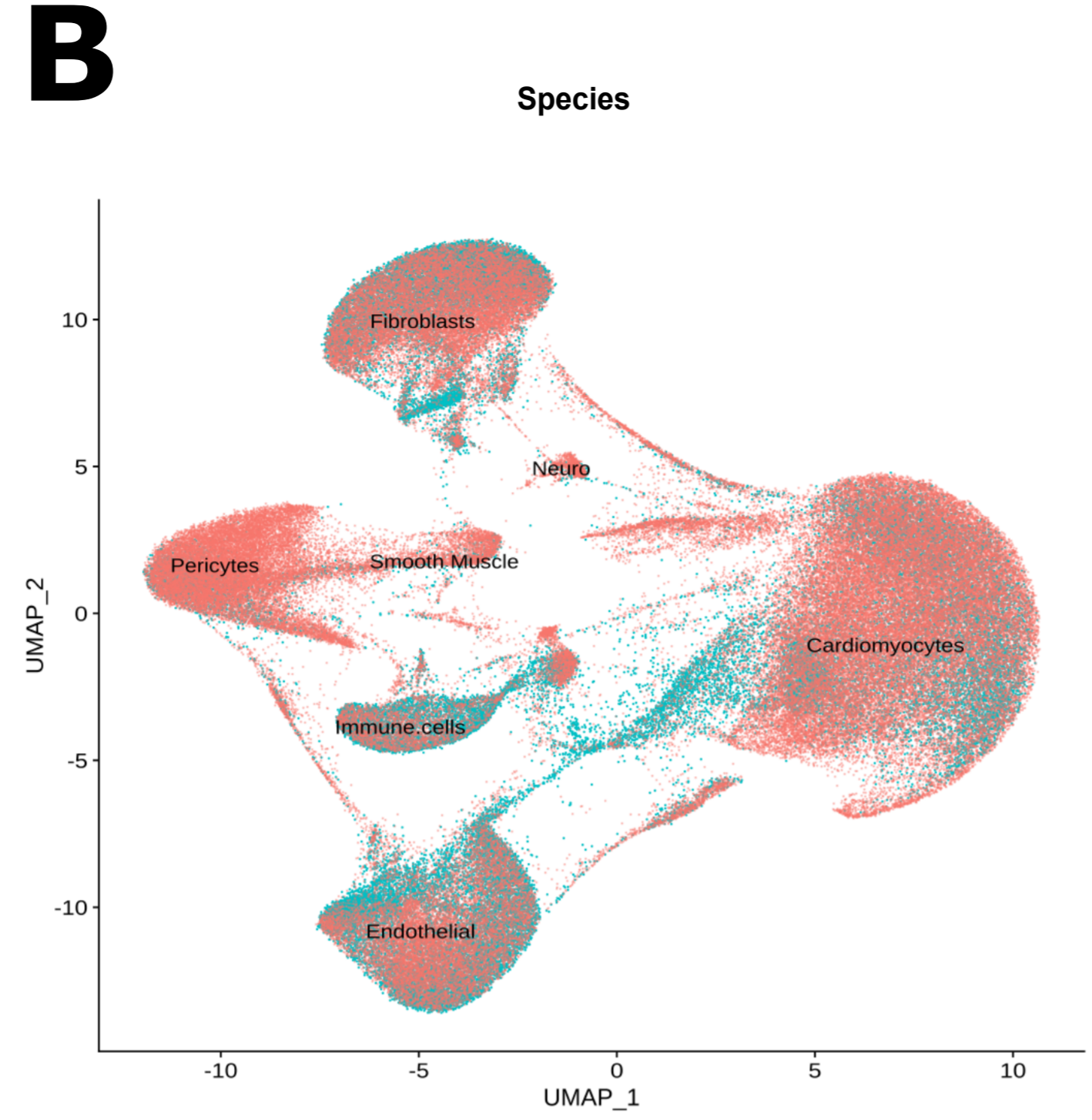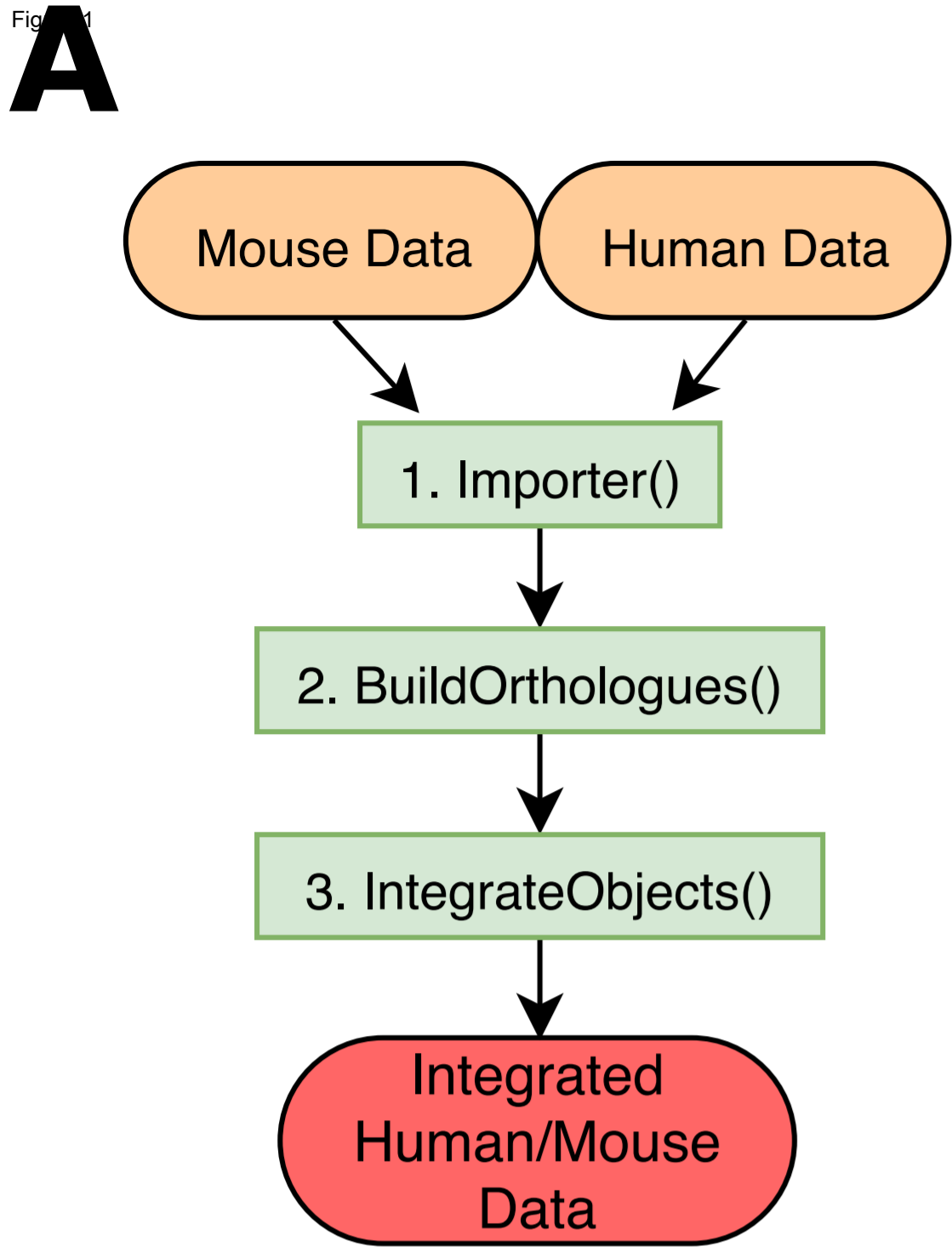
A
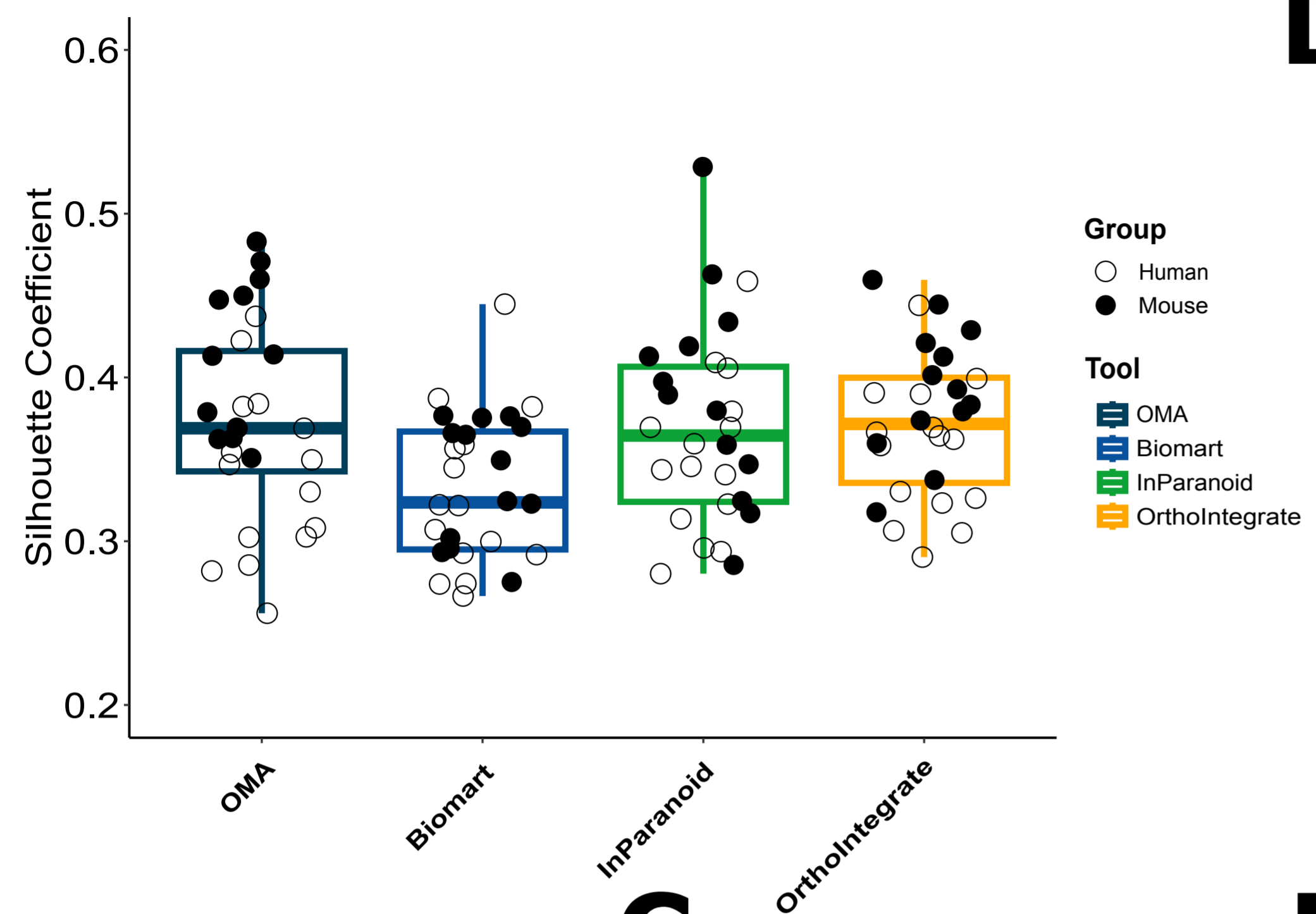
D

Tool	Batch Correction				Bio Conservation				Aggregate Score		
	GC	PCR	bASW	kBET	CCC	NMI	ILF1	SC	SMS	BSC	Total
OMA	1.0	0.368	0.649	0.839	0.882	0.831	0.923	0.373	0.714	0.752	0.733
Biomart	1.0	0.364	0.671	0.854	0.922	0.795	0.921	0.333	0.722	0.743	0.732
InParanoid	1.0	0.37	0.626	0.865	0.911	0.815	0.9	0.369	0.715	0.749	0.732
OrthoInt.	1.0	0.358	0.668	0.867	0.925	0.779	0.872	0.373	0.723	0.737	0.73

low   high
scaled per group

low   high
scaled per group

B

C

E

Enzyme Classification Conservation benchmark

Gene Ontology Conservation benchmark

Number of predicted pairwise orthologue relations

● OMA ● Biomart ● InParanoid ● OrthoIntegrate

F

Figure 3

Figure 4

Figure 5

Click here to access/download;Figure;Fig_5.pdf

A

Num. Terms p.adj 880 Total pathways clustered

Number of significant GSEA terms in each cluster (padj < 0.25)

mitochondrial oxidation
histone splicing metabolic methylation process
electron modification **deacetylation**

ribonucleoside metabolic nucleoside nucleotide atp proton
triphosphate **process** biosynthetic purine

actin permeabilization depolymerization filament cytoskeleton junction
organization assembly polymerization mitochondrial

cytoskeletal dense transepithelial vesicle
transport localization along transmembrane core vesiclemediated

growth regulation wnt receptor
signaling response **pathway** cellular stimulus

system angiogenesis epithelial embryonic glomerular
development differentiation morphogenesis cell guidance

sliding sensory filtration perception
contraction muscle skeletal cardiac heart uterine

others

p.adj
2.5×10^{-2}
2.5×10^{-1}
1

Similarity
0.6
0.4
0.2
0

Human HFrEF   Mice HFrEF

Human HFrEF   Mice HFrEF

B

20 most significant GO terms found in cluster 25 (padj < 0.25)

cell-cell adhesion
cell-substrate adhesion
glomerulus morphogenesis
extracellular matrix organization
regulation of phospholipase activity
skeletal system morphogenesis
substrate adhesion-dependent cell spreading
odontogenesis
enzyme-linked receptor protein signaling pathway
regulation of transmembrane transport
regulation of extracellular matrix organization
vascular process in circulatory system
potassium ion transmembrane transport
cellular response to amino acid stimulus
regulation of cell junction assembly
regulation of cell-substrate adhesion
regulation of chemotaxis

-log10(P)

C

20 most significant GO terms found in cluster 28 (padj < 0.25)

lymph vessel development
epithelial cell differentiation
embryonic pattern specification
protein localization to cell periphery
response to hormone
regulation of cell-cell adhesion
regulation of angiogenesis
cell population proliferation
positive regulation of protein localization to cell periphery
regulation of epithelial cell differentiation
positive regulation of canonical NF-kappaB signal transduction
positive regulation of monoatomic ion transport
response to retinoic acid
reactive oxygen species metabolic process
anatomical structure maturation
protein polyubiquitination
negative regulation of immune effector process
enzyme-linked receptor protein signaling pathway
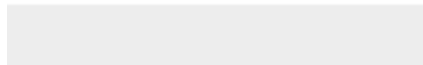
-log10(P)

Figure 6

Figure 6

Click here to access/download

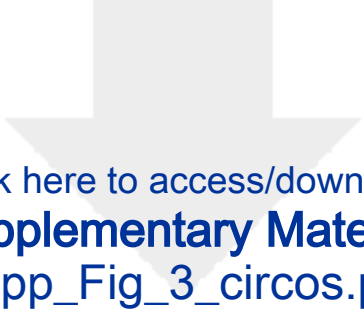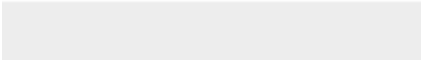**Supplementary Material**

Supp_Fig_1_pipeline.pdf

Click here to access/download
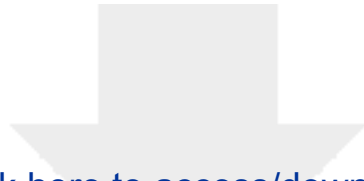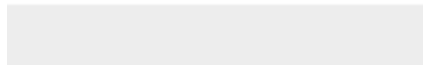**Supplementary Material**
Supp_Fig_2_UMAPs_comb.pdf

Click here to access/download
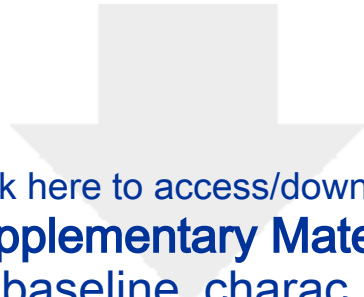**Supplementary Material**
Supp_Fig_3_circos.pdf

Click here to access/download
Supplementary Material
Supp_Fig_4_Upset_DEG.pdf

Click here to access/download
Supplementary Material
Supp_Fig_5_GSEA_Endothelial.pdf

Supplementary Figure 6

Click here to access/download
Supplementary Material
Supp_Fig_6_Hu_Mi_Ze.pdf

Supplementary Table 1

Click here to access/download
Supplementary Material
Supp_2_OrthologueAssignment_stats_mice.xlsx
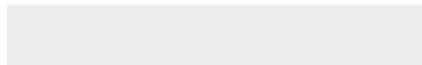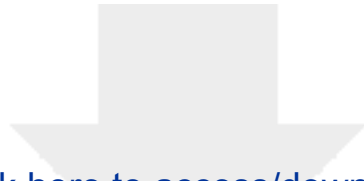
Click here to access/download
Supplementary Material
Supp_3_Human Mice DEG analysis HFrEF CTRL.xlsx
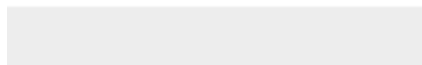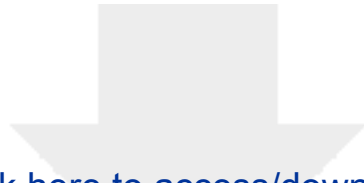
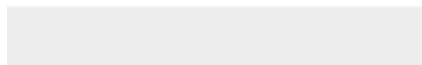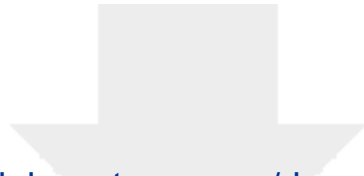Click here to access/download
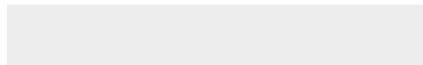**Supplementary Material**
Supp_4_Human Mice HFrEF GSEA CM.xlsx

Click here to access/download
Supplementary Material
Supp_5_Human Mice HFrEF GSEA EC.xlsx.xlsx

Click here to access/download
Supplementary Material
Supp_6_cell_numbers.xlsx

Fachbereich Medizin der
Johann Wolfgang Goethe-Universität
Frankfurt am Main

Institute of Cardiovascular Regeneration
Goethe-University Frankfurt am Main

University of Frankfurt ·Theodor-Stern-Kai 7 · 60590 Frankfurt

Dr. Edmunds
Editor in Chief
Gigascience

Institute of Cardiovascular Regeneration
Head of Bioinformatic Department:
Dr. David John
Tel.: +49 (0) 69 63 01-87952
E-Mail: john@med.uni-frankfurt.de
Director:
Prof. Dr. Stefanie Dimmeler
Tel.: +49 (0) 69 63 01-66 67
E-Mail: dimmeler@em.uni-frankfurt.de

Date: 2. February 2024

Submission of manuscript

Dear Dr. Edmunds,

Please find enclosed the revised manuscript entitled "**Improved integration of single cell transcriptome data demonstrates common and unique signatures of heart failure in mice and humans**", which we like resubmit to Gigascience for your consideration as a Research Article.

For the revision, we addressed all comments the reviewers made. As requested by the reviewers we succinct the biological part of the discussion and added a paragraph about long-non-coding RNAs. Additionally, we integrated details about the isolation and preparation of the sequencing library.

The review process finalized the paper quality and therefore we would like to thank the reviewers for their remarks and suggestions.

We believe that our manuscript provides an advanced approach to test the comparability of disease models in mice versus humans, or other species. These insight might lead to the improvement of animal models to test novel therapeutic interventions and in addition may **provide novel insights into so far unexplored human specific disease signatures**.

We hope that the revised manuscript might be considered for external reviewing and be of interest for the readers of Gigascience.

Thank you very much for your generous consideration.

Yours sincerely,

Dr. David John