Master's degree project

A multi-organ single-cell transcriptomic map of the pig

Submitted by
Emilio Skarwan

Supervised by
Dr. Linn Fagerberg

15th of May, 2023
MSc Molecular Techniques in Life Science

Science for Life Laboratories
Karolinska Institutet
Stockholm Universitet
KTH Royal Institute of Technology

Index

Abstract

Biology has a long history of drawing systematic maps of diversity. Scientific and technological advancements enable now to identify and describe the cellular diversity of cells based on gene expression through single cell transcriptomics.

In this report, I discuss my work involved in the generation of a single cell atlas of the pig based on 11 tissues and 9 brain regions. I report my approach to data processing and cell type annotation which lead to the generation of expression profiles of 66 cell types of the pig, based on transcript data of 186,247 cells. Additionally, I annotated all protein coding genes in terms of their specificity to cell types, which allows to establish relationships between cell types and tissues based on gene expression as well as efficient exploration of the atlas. Rigorous comparisons between pig cell types and bulk tissues as well as human cell types, indicate that the atlas' data is consistent across datasets, however, also highlights the need for ambient RNA data corrections procedures before final publication.

This atlas is developed as a complement to bulk tissue data of the pig available at www.rnaatlas.org and should prove useful as a future reference tool for fields in biomedical and pharmaceutical research, where pigs are established animal model, as well as in agricultural research and research and control of zoonotic diseases.

Introduction

With the widespread acceptance of cell theory in the mid 19th century and the consensus among scientists acknowledging the cell as the fundamental and indivisible unit of life, biologists have been creating maps of cellular diversity based on their morphology and function. A well-known example is Ramón y Cajal's map of cells of the nervous system[1]. Advances in the 20th century led to prove DNA as carrier of genetic information,[2] which paved the way to determine the rules to how genetic information is stored and how it flows within biological systems, as proposed through Crick's sequence hypothesis and central dogma.[3] By the turn of century, in 2001, scientists of the Human Genome Project had published most of the human genome's sequence,[4,5] which profoundly changed molecular biology research. However, this also led to the next project proposition: Sidney Brenner's *CellMap* in 2002: defining cells based on the genes they express and by this create a map of all the cells in an organism, and of molecules in a cell.[6]

Research in the last decade have followed Brenner's proposition to the letter. Sequencing technologies advanced so that 2009, the first transcriptome of a single cell was sequenced.[7] Early work by Sten Linnarsson, et al. in 2011 pioneered the field by characterising 85 single cells based on their gene expression through RNA sequencing.[8] Since then, research and innovation of single cell -omics technologies aided by computational advancements have allowed the development of large scale multi-organ *CellMaps* or cell atlases by consortia such as the Human Protein Atlas,[9] and the Tabula Sapiens[10] each composed out of thousands of cells per organ.

Although other methods exist to characterize gene expression in single cells,[11] single cell transcriptomics has established itself as a routine method to study the gene expression activity of single cells.[12] This is due to lower costs, simpler data analysis and high throughput while identifying and quantifying RNA transcripts with high sensitivity.[13]

Cell characterisation through picturing a cell's RNA expression profile has enabled us to better understand how a diversity of cells collaborate to enable tissue and organ function.[14] It has permitted advances in cell development by discovering novel cell states and cell types, and at the same time it has enabled disease mechanisms to be characterised by cell malfunction.[14] Similarly, it has enabled to uncover protein functions and establish reaction networks based on gene co-expression across cell types and tissues.[14] Consequently, this changed the approach of drug development by introducing novel ways to identify drug targets.[14] Single cell transcriptomics permits to understand organ system function and dysfunction through the lens of a cell.

As biomedicine and drug development research become more focused on studying the function of the single cell as a system, so does the need to describe the cells of model organisms. A baseline reference cell atlas of model organisms is essential for establishing links between organisms at a cellular level and assess the potential of model organisms to model a certain disease or to trial for a potential treatment. Additionally, thorough cell atlases of livestock animals can streamline research for the prevention and control of novel zoonotic diseases by mapping the presence of possible viral entry factors.[15]

With this necessity in mind, the Human Protein Atlas (HPA) has proposed itself to assemble a mammalian RNA atlas (www.rnaatlas.org), which in the future aims to include both bulk and single cell gene expression data from mammalian model organisms. The current version, however, only includes the bulk data from the pig (*Sus scrofa domesticus*) from 98 tissues.[16]

The pig is today an established model organism used in biomedical and pharmacological research.[17] Their applications range a widely from drug development, vaccine testing, and genome editing to research in cardiovascular, dermatological, developmental, neurological, or respiratory disorders as well as various cancer types, among other things.[17] In spite of showing to be more challenging to handle than smaller model organisms such as rodents, pigs have higher similarities to humans with respect to size, anatomy and physiology as well as in their immunology and tissue function.[18]

Herein, I report my work on the processing, assembly, and analysis of a single cell RNA atlas of the pig spanning 11 distinct organs and 9 brain regions based on data from Wang et al.[19], Zhu et al.[20] Zhang et al.[21] This new pig cell atlas is built upon the expression of 186,247 cells, aggregated into 66 unique consensus cell types. Throughout my project I aimed to: (1) investigate upon cellular composition of each tissue through manual cell cluster annotation to establish a baseline cell type expression for the atlas; (2) annotate all protein coding genes of the pig based on cell types specificity and distribution to establish a navigable framework for the atlas and enable data exploration; (3) compare the pig single cell data to the bulk tissue transcriptomes and to human cell types transcriptomes to assess the quality and consistency of the data and annotations.

The pig single cell atlas should become easily comprehensible and accessible at the HPA's mammalian RNA atlas (www.rnaatlas.org) and complements the already available bulk RNA atlas.

Methods

Single cell and single nuclei transcriptomic datasets

The pig single cell atlas integrates data published and generated by three previous studies[19–21] all based on tissue sections collected from three-way hybrid of Landrace, Large White and Duroc pigs (*Sus scrofa domesticus*). In total, 24 transcript libraries stemming from 11 tissues and 9 brain regions were integrated. (**Figure 1A**)

The data used published by Wang et al.[19] was downloaded from the CNGB Sequence Archive of China National GeneBan Database (CNGBdb) under accession "CNP0002165". Their data provided me with both single cell RNA sequencing (scRNA-seq) and single nuclei RNA sequencing (snRNA-seq) datasets. For the generation of the scRNA-seq libraries they sectioned liver, spleen, retina, brain, lung, visceral adipose, subcutaneous adipose and intestine tissues and isolated the PBMC (peripheral blood mononuclear cells) from a six-month-old healthy pig. The brain section consisted of seven 0.5 g sections of the neocortex (cerebral cortex), cerebellar cortex (cerebellum), caudate nucleus (basal ganglia), thalamus, hypothalamus, hippocampus, hypothalamus, and pons. They constructed the library using Single Cell 3' Gel Bead and Library kit v3 from 10 Genomics, converted the library using MGIEasy Universal DNA prepration reaction kit (BGI) and sequenced using the DNBSEQ-T7 platform (MGI).

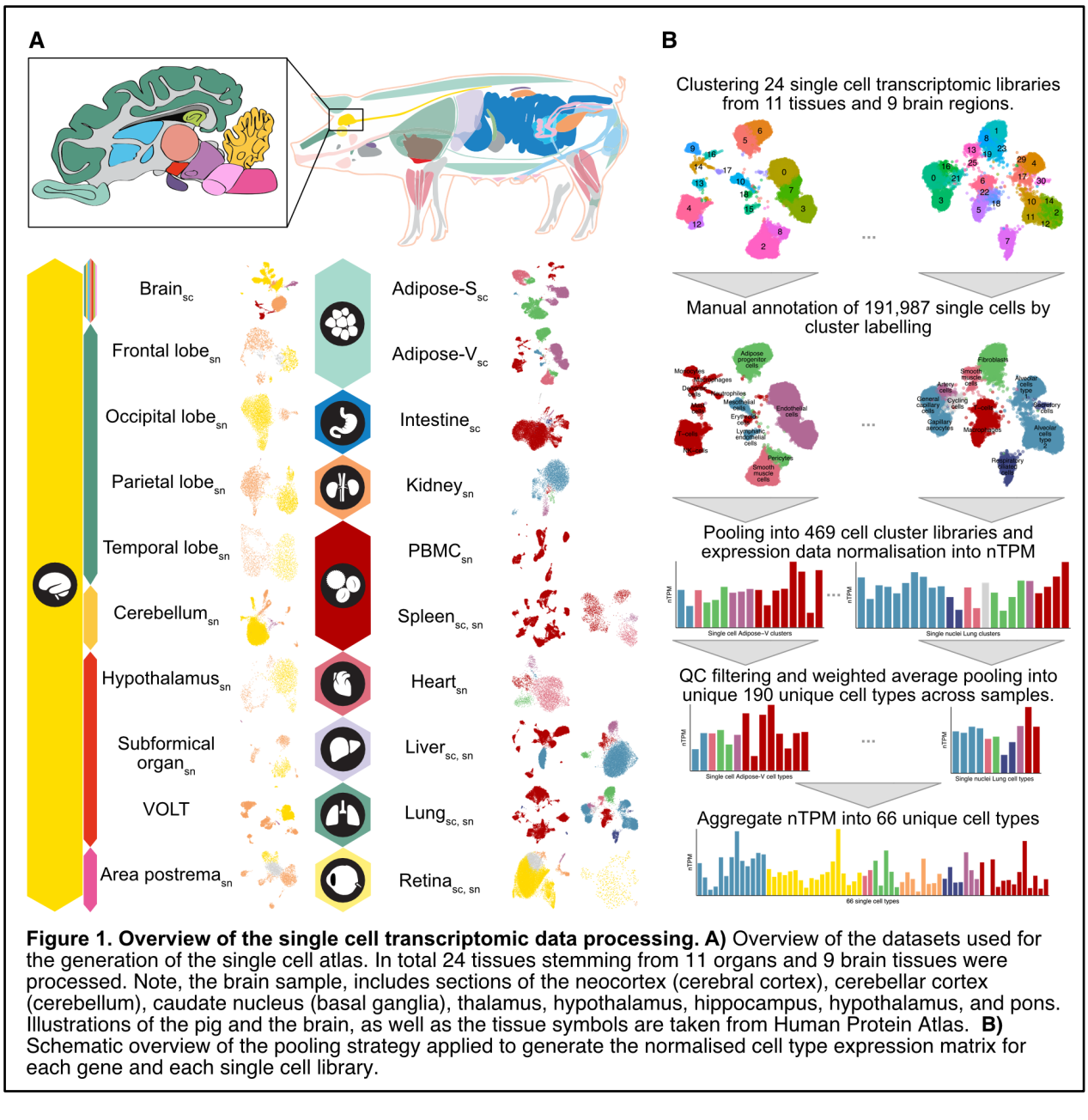


Figure 1. Overview of the single cell transcriptomic data processing. A) Overview of the datasets used for the generation of the single cell atlas. In total 24 tissues stemming from 11 organs and 9 brain tissues were processed. Note, the brain sample, includes sections of the neocortex (cerebral cortex), cerebellar cortex (cerebellum), caudate nucleus (basal ganglia), thalamus, hypothalamus, hippocampus, hypothalamus, and pons. Illustrations of the pig and the brain, as well as the tissue symbols are taken from Human Protein Atlas. **B)** Schematic overview of the pooling strategy applied to generate the normalised cell type expression matrix for each gene and each single cell library.

The snRNA-seq datasets include transcriptomic libraries from the heart, kidney, spleen, liver, retina and four brain regions: cerebellum, subfornical organ, vascular organ of lamina terminalis (VOLT) and area postrema. These samples were sectioned from a three-month-old healthy pig. They implemented the MGI DNBelab C series reagent kit (MGI) for library construction and sequenced using the DNBSEQ-T7 platform.

The data published by Zhu et al.[20] was downloaded from CNGBdb under accession CNP0000686. They provided with snRNA-seq data from five brain regions: frontal lobe, parietal lobe, temporal lobe, occipital lobe, and hypothalamus stemming from a 3-month-old healthy pig. They constructed the libraries using Chromium Single Cell 3' Reagent Kits v2 (10x genomics), performed library conversion using MGIEasy Universal DNA preparation reagent kit (BGI) and sequenced using BGISEQ-500.

The snRNA-seq data set of the Lung was published by Zhang et al.[21] was accessed through CNGBdb under accession number CNP0001486. They dissected eight lung tissue pieces from

three three-month-old (male) healthy pigs. Library construction was preformed based on the MGI DNBelab C series reagent kit (MGI) and sequenced using DNBSEQ-G400 and DIPSEQ-T1 from MGI.

The sampling of all tissues above were reported as approved by their responsible ethics committees.

Preprocessing of single cell data and cell barcode clustering

Fastq files were aligned using cellranger 6.1.2 to the *Sus scrofa* reference from Ensemble build version 109 based on the genome assembly Sscrofa 11.1 (GCA_000003025.6)[22]. I input the filtered expression data into Scanpy (version 1.9.1) for downstream analysis running under Python version 3.9.5. In Scanpy, every sample underwent doublet filtering using scrublet (version 0.2.2) set to an expected doublet rate of 0.1. Outliers regarding percentage of mitochondrial genes detected per cell barcode were defined by median absolute deviation (MAD) thresholding, where X_i is the percentage of mitochondrial genes in a barcode:

$$MAD = median(|X_i - median(X)|)$$

Cells accounting for more than three MADs differences in percentage of mitochondrial genes compared to all barcodes in the sample, were considered outliers. Mitochondrial genes, were defined as all the transcripts from the mitochondrial chromosome, as described in the Ensemble annotation.

Similarly, cell outliers were identified if they had over 5 MADs of difference on either: the natural logarithm (log_e) of 1 + the total read counts, on log_e of 1 + the number of genes detected, or on the percentage of the top 20 most expressed genes in the sample library. Outliers were not included in subsequent analyses. Genes detected in less than 20 cell barcodes were filtered out as were cell barcodes containing less than 200 genes.

Subsequently the count data was normalized to have a total of 10,000 counts per cell and underwent log1p scaling ($\log_e(1 + x)$). Highly variable genes were detected using the highly_variable_genes function in scanpy on standard setting. Before cell clustering, the effect of mitochondrial genes was regressed out, PCA, neighbourhood graph and UMAP were computed under default parameters based on the previously determined highly variable genes. Clusters were defined through the Leiden algorithm.

Preprocessing of single nuclei data

For the alignment, I implemented STARsolo under the solotype setting for CB_UMI_Simple. Additional settings were specified to 1MM_multi_Nbase_pseudocounts for soloCBmatchWLtype, MultiGeneUMI_CR for soloUMIfiltering, and 1MM_CR soloUMIdedup. Depending on library preparation method (MGIEasy Universal or 10x Chromium Single cell v2), the barcode length and barcode whitelisting were adjusted to match the library.

In scanpy, single nuclei fastq files stemming from the matching tissues were pooled together. The expected doublet rate in scrublet was set to 0.05 for the MGIEasy libraries. Otherwise, outlier detection and filtering followed the same rules as described above, as did the dimensionality reduction and clustering. Since the lung snRNA-seq sample, was the only one, which integrated

sections from multiple organisms, only this sample went through batch corrections through scanpy's implementation of harmony[23] (harmony_integrate).

To note, selected snRNA-seq samples (hypothalamus, VOLT, parietal lobe, frontal lobe, area postrema, and subformical organ) underwent re-clustering with manual settings regarding the calculation of variable genes, dimensionality reduction and clustering, due to difficulties in finding a clustering pipeline that would work with all samples equally as well as rather a low cell count in the sample.

Cell type cluster annotation

I manually annotated the resulting 469 cell clusters from the 24 libraries using scanpy. The cell transcript markers applied for cell identification were based on pig orthologs of the markers applied by the Human Protein Atlas, as well as cell markers used by Wang et al.[19] Additional markers were included through extensive literature research and used by other organ atlases. Cell type clusters that resulted in mixed cell types were not used for further analysis.

Generation of cell type pseudo-bulk expression dataset and normalisation:

After annotation of the 469 clusters, a pseudo-bulk expression dataset was generated by summing up the gene counts for every cell barcode inside the same cluster. Only protein coding genes, as defined by the Human Protein Atlas, were included into the pseudo-bulk. This reduced the expression matrix to 22,063 protein coding genes, for every cell cluster in a sample.

The expression count matrix of every cluster went through TPM normalisation, to form the protein coding TPM expression data (pTPM):

$$RPK = \frac{gene\ count}{transcript\ length} \qquad TPM = 10^6 \frac{RPK}{\sum(RPK)}$$

The pTPM expression data was normalized using the Trimmed mean of M values procedure (TMM).[24] To this I refer to as normalized pTPM values (nTPM). Matching cell types inside each sample were pooled together through calculating the weighted average, which was weighted by the cluster cell count. Cell clusters showing mixed cell types as well as low confidence annotations were not included into the weighted average expression matrix. This reduced the number of cells effectively used to generate the atlas to 186,247.

Finally, the expression profile for each cell type was calculated through the unweighted average gene expression of matching cell types across samples. This results in an expression profile of 66 cell types. For clarity, **Figure 1B** depicts the processing steps after the cell clustering.

Gene specificity and distribution classification and Tau score

The gene expression profiles of all protein coding genes were annotated with categories describing specificity and distribution of the genes towards the cell types or tissues. The definition to the specificity and distribution categories are listed in the tables below, every protein coding gene is assigned one specificity category and one distribution category. Cell type enriched, group enriched, or cell type enhanced genes are collectively called cell type elevated genes.

Specificity category	Definition
Cell type enriched	4-fold higher nTPM expression in one cell type compared to any other cell type.
Group enriched	4-fold higher mean nTPM expression of a group of 2-10 cell types (or 2-5 tissues) compared to the maximum of the remaining of cell types.
Cell type enhanced	4-fold higher mean nTPM in a group (1-10 cell types or 1-5 tissues) compared to the mean expression.
Low cell type specificity	nTPM ≥ 1 and not in any of the categories above.
Not detected	nTPM < 1 in all cell types.

Distribution category	Definition
Detected in single	nTPM ≥ 1 in a single cell type.
Detected in some	nTPM ≥ 1 in under 31% of cell types.
Detected in many	nTPM ≥ 1 in 31% of cell types or more.
Detected in all	nTPM > 1 in all cell types.
Not detected	nTPM < 1 in all cell types.

As a complement to the gene categorisation, a Tau score (τ) or tissue specificity index was calculated using $\log_{10}(x+1)$ transformed. For every gene τ is calculated as defined by Yanai, et al, where N is the number of cell types and x_i is the expression value relative to the highest expression of gene i.[25]

$$\tau = \frac{\sum_{i=0}^{N}(1-x_i)}{N-1}$$

The annotations used for Figures 4, 5, 6, 7, S6 were carried out based on 65 cell types of the pig, that is excluding platelets. The annotations used for Figures 6 were based on bulk grouped tissue data[16] of the tissues listed on Figure 6B. The annotations used for figure S7, were based on cell type transcript data only from brain tissues and retina tissues and bulk region tissue data[16] from the tissues listed in the figure. For Figure 7, the annotations of the bulk tissues were based on the whole region tissue dataset of the pig[16]. For Figure 8, the annotations were computed based on a consensus dataset aggregated as described in Figure S8 from the human single cell data[9] and the pig single cell data. Computed only including pig-human orthologs genes, defined by the human protein atlas.

Dimensionality reduction

Using the R package pcaMethods, the principal components (PCs) for were calculated based on $\log_{10}(x+1)$ transformed and center scaled values of nTPM expression data. PC1 and PC2 were used for principal component analysis (PCA) visualisation.

To perform UMAP visualisations, the principal components were calculated as described above. Principal components accounting for 80% of variability were selected to perform UMAP dimensionality reduction based on the uwot library in R.

Correlation calculations

Using the library stats, the function cor was called to calculate correlation between expression profiles, where 'spearman' was chosen as a method. Spearman distance is 1 - spearman correlation. The output was used to construct dendrograms and clustered heatmaps. Dendrograms were constructed using the hclust function from stats in R to run complete-linkage hierarchical clustering analysis on the dissimilarities. Clustered heatmaps were constructed using the pheatmap function from the pheatmap package based on the cross cell type correlation. Complete linkage method was applied for clustering.

To construct the spearman network plot seen in figure 7, the pairwise spearman correlation was computed between single cells and bulk region tissues based on enriched gene using the pairwise_cor function of the widyr package. Enriched genes are genes annotated as cell type or tissue enriched or group enriched. After excluding any correlation under 0.60, the top 3 cell type to tissue and tissue to cell type correlations were taken to build the network.

Hypergeometric tests

To compare the overlap of enhanced genes between two datasets, a hypergeometric test was performed. For this, both datasets being compared consisted of the same genes or solely of matching homolog genes. Both datasets were then categorized in terms of specificity and distribution. Applying the function phyper from the stats package in R, a hypergeometric test was performed. Here, q was defined as the total matching elevated genes in both datasets; m the lowest number elevated genes between the two tissues being compared; n was the total number of genes or homolog genes subtracted by m, k was the total number of orthologs elevated in either tissue. FDR was computed from the resulting p value for multiple comparisons applying the Benjamini Hochberg p adjustment method under the p.adjust function of the stats package in R.

Bulk pig tissue data

Pre-release data of the bulk tissue transcriptomic data from the pig RNA atlas (www.rnaatlas.org) based on ensemble build 109 were used for the comparison.[16]

Human single cell type data

For comparisons between pig and human, single cell type atlas of the Human Protein Atlas[9] (www.proteinatlas.org/humanproteome/single+cell+type) was used, based on Ensemble 103. Batch correction using the removeBatchEffect in limma was applied before computing the UMAP shown in Figure 8C.

Data downstream analysis and visualisation

R was used for all analysis following the pTPM computation. R version 4.2.2 (2022-10-31) was ran through rStudio Version 2022.12.0+353. Most visualisations were first plotted using ggplot2 (3.4.1) and edited for aesthetic and labelling purposes in affinity designer 2 (v. 2.0.4). Other packages used through R are: geomtextpath (v. 0.1.1), ggraph (v. 2.1.0.9000), edgeR (v. 3.38.4), limma (v. 3.52.4),

ggrepel (v. 0.9.3), patchwork (v. 1.1.2), pheatmap (v. 1.0.12), ggplotify (v. 0.1.0), plotly (v. 4.10.1), uwot (v. 0.1.14.9000), Matrix (v. 1.5-3), ggdendro (v. 0.1.23), ggalluvial (v. 0.12.5), lubridate (v. 1.9.2), forcats (v. 1.0.0), stringr (v. 1.5.0), dplyr (v. 1.1.0), purr (v. 1.0.1), readr (v. 2.1.4), tidyr (v. 1.3.0), tibble (v. 3.2.1), tidyverse (v. 2.0.0), pcaMethods (v. 1.88.0), Biobase (v. 2.56.0), BiocGenerics (v. 0.42.0), biomaRt (v. 2.52.0).

Code availability

Code applied for acquisition of data, alignment, data processing and figure visualisation is accessible in github under github.com/emiliosk/Pig_sc_atlas.

Results

Annotating and generating baseline pig single cell atlas data

As depicted in **Figure 1B**, I manually annotated 469 cell clusters or 191,987 cells coming from 24 libraries. For this I referenced orthologs of cell types of markers by the Human Protein Atlas (HPA), markers used by Wang et al.[19] and orthologous of markers used in single cell atlases of the human and mouse of specific organs. The resulting annotated UMAP plots of every transcriptomic library can be found in **Figure S1** and **Figure S2.**

With the annotation in place, the expression profiles of each cell inside a cluster were pooled together and TMM normalized to form a pseudo-bulk expression matrix. To control for correct normalization and overall relationships between the clusters, a PCA and a UMAP plots were calculated. The UMAP and PCA plots (**Figure 2A, 2B**) show that the clusters cluster based on cell type, on tissue type as well as on library preparation method, i.e. snRNA-seq or scRNA-seq.

Due to differences in resolution across libraries, a global cell type name had to be established for all cell cluster names. (**Figure S3**). This resulted in a total of 66 cell types detected. After quality control of the clusters, I integrated the cluster pseudo-bulk data to form a matrix containing the expression profiles of the 66 unique cell types. This matrix constitutes the baseline expression data for the atlas and was used for all further analysis. **Figure 2C** shows in a dendrogram the relationships between cell types and makes visible, how related cell types cluster together. Similarly, a clustered heatmap (**Figure S4**) depict detailed insights on the similarities and dissimilarities between expression profiles of all cell types.

Figure 3 summarizes the population of cell types detected in each library presented in Figure S1 and S2. It is noticeable that there is a wide variation in both number of cells in cell types and number of cells in a sample. The sample with lowest cell count (retina_snRNA-seq) amounts to only 424 cells, while the sample with the highest cell count amounts to 16,452 cells (spleen_scRNA-seq). Generally, scRNA-seq libraries yielded higher number of cells. The cell type with highest cell count on the other hand were T cells (N = 27,811) followed by B-cells (N = 13,481) and hepatocytes (N = 12,133). Cell types with the lowest population count were megakaryocytes (N = 16). **Figure S5** presents the relative cell count for each tissue.

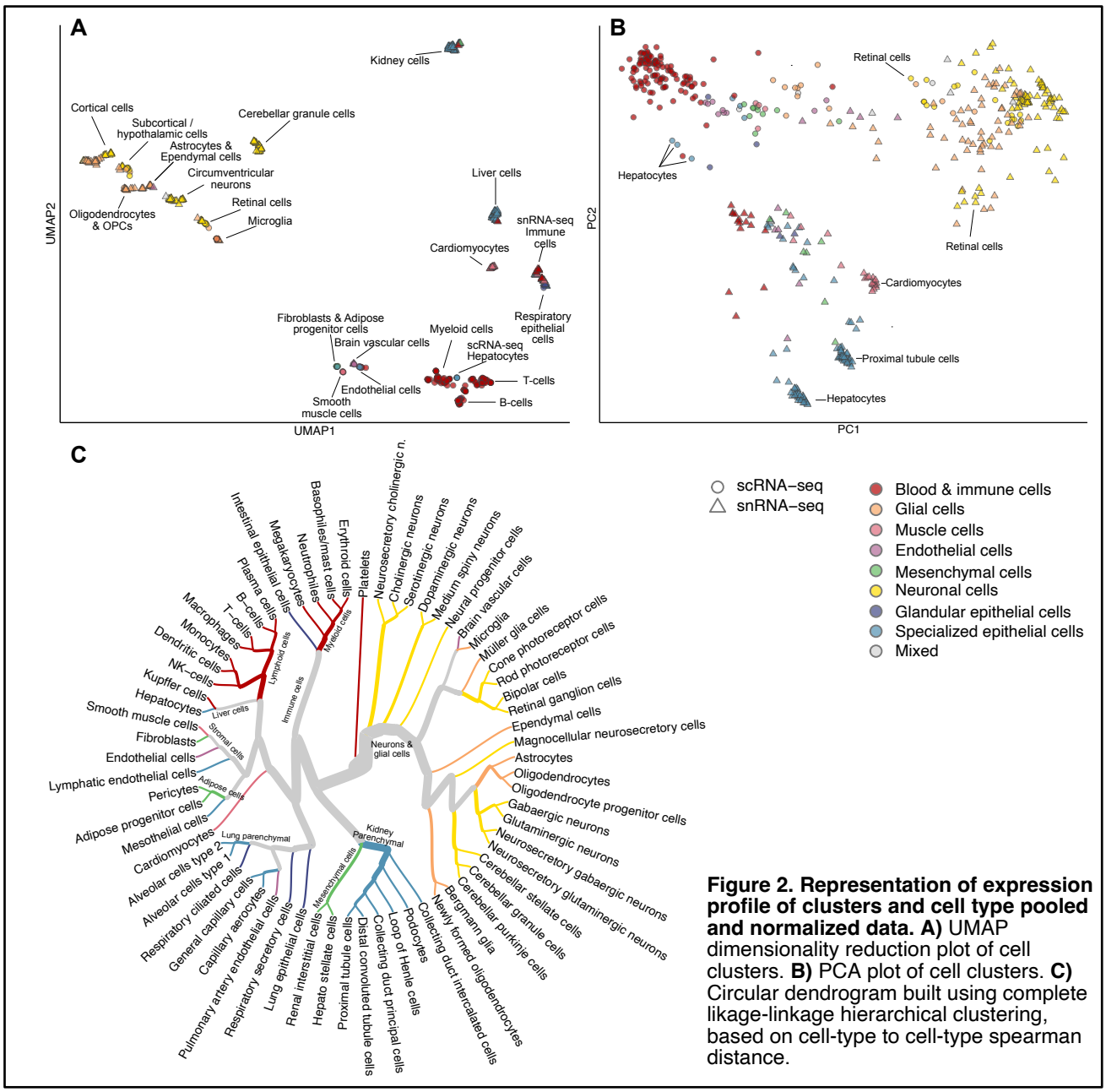


Figure 2. Representation of expression profile of clusters and cell type pooled and normalized data. A) UMAP dimensionality reduction plot of cell clusters. **B)** PCA plot of cell clusters. **C)** Circular dendrogram built using complete likage-linkage hierarchical clustering, based on cell-type to cell-type spearman distance.

Ultimately, to exemplify the practicalities of establishing an atlas based on pseudo-bulked cell type data, I programmed an atlas enquiry interface. I present in **Figure 9** examples of 3 of the 22,063 protein coding genes possible to enquire through this atlas.

Protein coding gene annotation

Having the pooled expression matrix in place for every cell, I proceeded with the whole genome annotation of protein coding genes of the pig. Every protein coding gene got assigned a category with regards to cell type distribution and cell type specificity. (**Figure 4**)

Briefly, the categories with respect to distribution are detected in single, detected in some, detected in many, detected in all, and not detected. These categorizations make apparent, that from the 22,063 protein coding genes of the pig, 1,643 or 7.4% of genes were not detected. Similarly, 2,440 or 11.1 % of genes got detected in all cells. And 1070 or 4.8% of genes were detected in a single cell type. (**Figure 4A**) The cell type detected with the lowest number genes were platelets (N = 2,756), however due to their low number of genes, they were not included in the annotations.

Figure 3. Cell counts identified for each cell type by sample. A total of 66 cell types were identified across all 24 libraries. Relative cell count was calculated as the quotient of the counts per cell type in a sample and the total cell count of a sample. The barplots depict the total cell count in each sample (top) and each cell type (right).

Otherwise, neurosecretory cholinergic neurons were the cell type with fewest genes, with 6,929 genes. The cell type with the highest number of detected genes were alveolar type 2 cells with 15,290. The average number of genes detected in a cell type is 12,293.15 genes. (**Figure S6A**) Alveolar type 2 cells had highest number of genes detected in a single a cell type (N = 107), followed by the general capillary cells (N = 90) and cardiomyocytes (N = 70). (**Figure S6B**).

The categories assigned to each gene regarding cell type specificity are, from highest to lowest specificity: cell type enriched, group enriched, cell type enhanced and low cell type specificity. Only 18.5% of genes, or 4,088 genes show low cell type specificity. Most genes (12,497 or 56.6%) however show to be categorized as tissue enhanced. Then 8.1% or 1,777 genes are group enriched and 9.3% or 2,058 genes are tissue enriched. The number of elevated genes per cell type vary

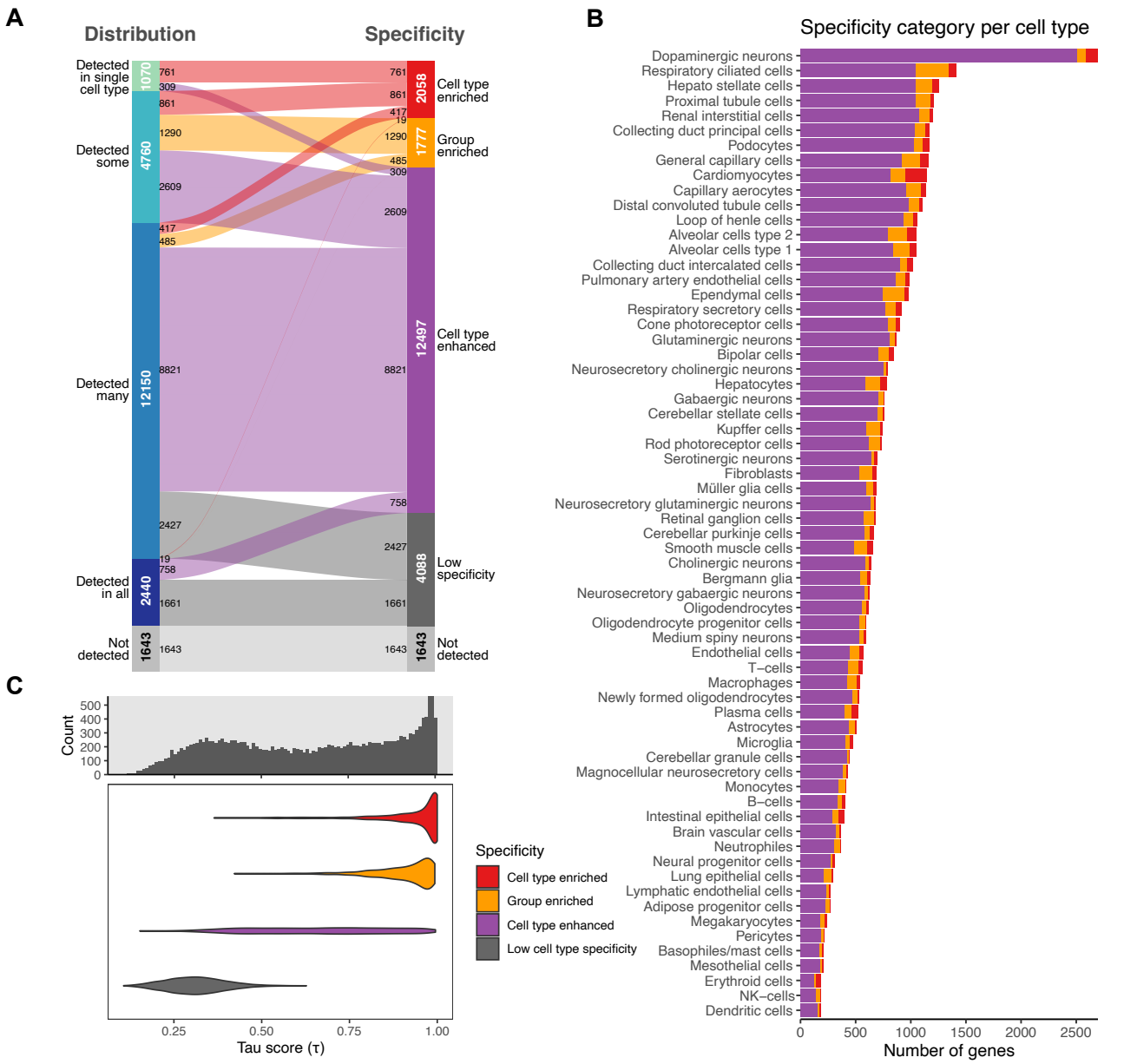


Figure 4. Distribution, Specificity and Tau (τ) specificity index annotation of protein coding genes. A) Alluvial plot depicting the number of assigned genes for each of the five distribution categories and five specificity categories, as well as the number of the number of genes in every specificity and distribution category combination. **B)** Bar plot showing the number cell type elevated genes (cell type enriched, group enriched and cell type enhanced) per cell type. **C)** Bottom, a violin plot depicting the tau scores for each gene inside split by specificity category. Top, a histogram depicting the total distribution of tau scores in the single cell dataset.

between 184 for dendritic cells and NK cells, goes up to 1,417 for respiratory ciliated cells and 2,697 for dopaminergic neurons. (**Figure 4B**). More in detail, the tissue with lowest tissue cell type enriched genes (in red) are neutrophiles and pericytes with only two cell type enriched genes each. While the top 3 cell types with highest tissue enriched genes are Dopaminergic neurons with 109 and cardiomyocytes with 196. More specifically, these dopaminergic neurons, are hypothalamic domain neurons, no other exclusively dopaminergic neurons were detected.

As a complement to the discrete category annotation of genes, a numerical specificity index Tau (τ) was also added to the annotation of every gene. As we see in **Figure 4C** a higher specificity category is usually accompanied with a high Tau score, the opposite is true as well.

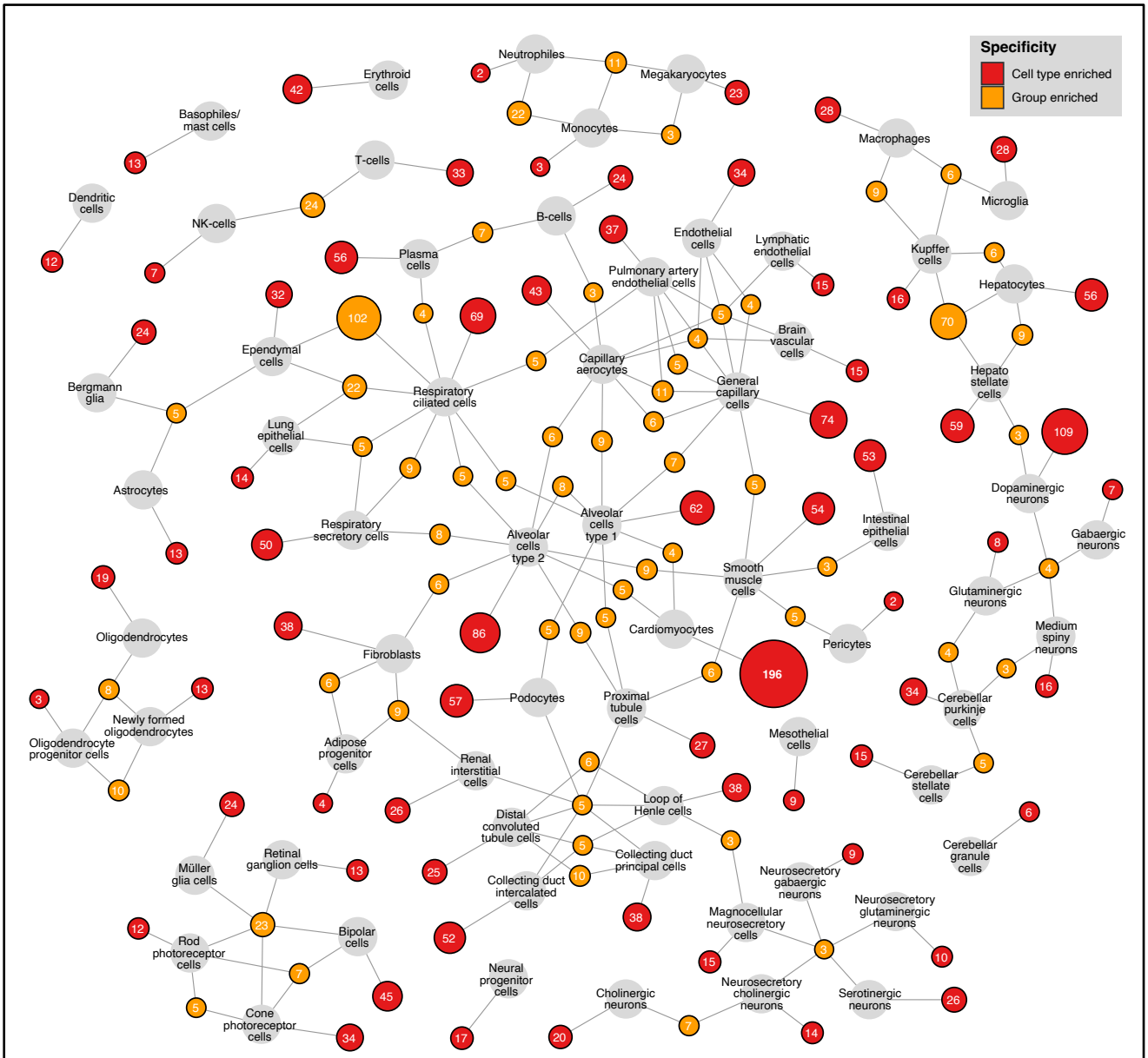


Figure 5. Network plot build upon shared group enriched genes between cell types. Additionally, for each cell type a node labelled with the number of cell type enriched genes is assigned. To reduce visual complexity, depicted are only group enriched node containing at least 5 genes or group enriched nodes which ranked top 2 for a given cell type provided they accounted with 3 or more group enriched genes.

To demonstrate the utility behind the gene annotation in the explorative navigation of the single cell atlas I computed a network plot. (**Figure 5**) The plot shows for every cell type the number of assigned cell type enriched genes in red, and the shared group enriched genes shared between cell types in orange. This plots highlights, that ependymal cells in the brain and respiratory ciliated cells in the lung share the highest amount of group enriched genes (N = 102). Similarly, the cells in the liver hepato stellate cells, Hepatocytes and Kupffer cells share 70 group enriched genes. These shared enriched genes between cell types can subsequently be explored in detail inside the atlas as exemplified in **Figure 9C.**

Comparison to pig bulk tissue data

To assess the quality and consistency of the data, I proceeded to compare the transcriptomic profiles of the cell types with the profiles of bulk tissues of the pig. For the first part of the comparison, I compared the single cell type data with the transcript data of the matching tissues of

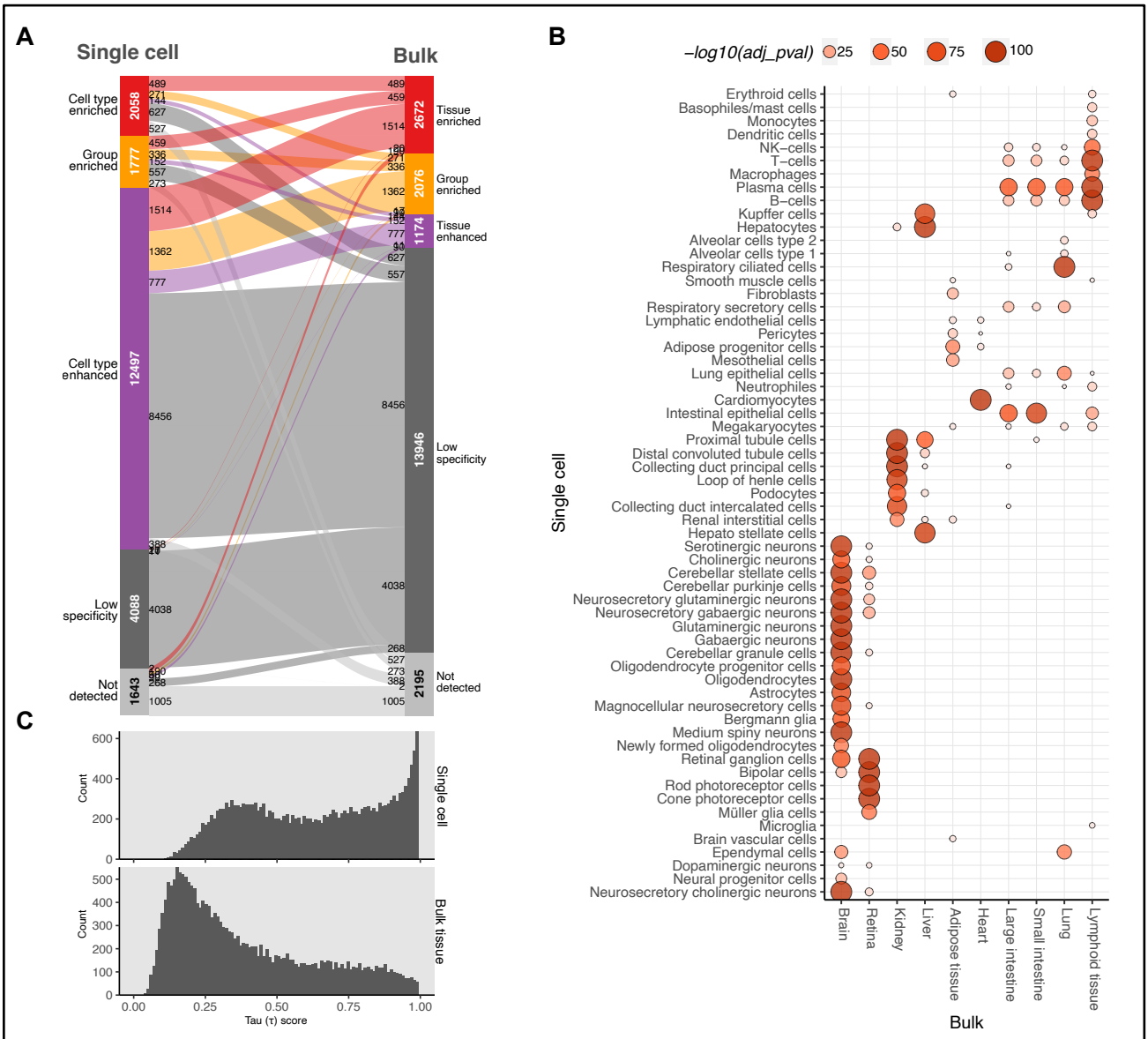


Figure 6. Comparison between bulk transcriptomic data and single cell pig data. A) Alluvial plot comparing the assigned specificity categories based on bulk data and single cell data. Note that the classification for the bulk data was calculated solely on data composed of brain, retina, kidney, liver, adipose tissue, large intestine, lung and lymphoid tissue data. **B)** Hypergeometric test, based on matching elevated genes between tissues and cell types. **C)** Comparison between Tau scores computed using both datasets.

origin. In other words, I compare the single cell type data with the transcriptomes of the brain, retina, kidney liver, adipose tissue, heart, large intestine, small intestine, lung, and lymphoid tissue. However unavailable were the transcriptomic bulk data of the PBMC.

As the core for my comparison, I annotated again all genes in terms of specificity for the bulk tissue dataset based solely on the expression profile of the tissues mentioned above. (**Figure 6A**) There are two things immediately visible after comparing the annotations of genes based on single cell with the annotations of genes based on the bulk data. Firstly, there are more genes detected in the single cell dataset. Secondly, there is a higher amount of tissue enhanced genes in single cell data, compared to bulk data. In the bulk data there is a predominant number of genes with low tissue specificity.

I also compared the Tau score calculated for each gene based on the two different datasets. The histograms in **Figure 6C** makes noticeable that the single cell dataset computes for higher number

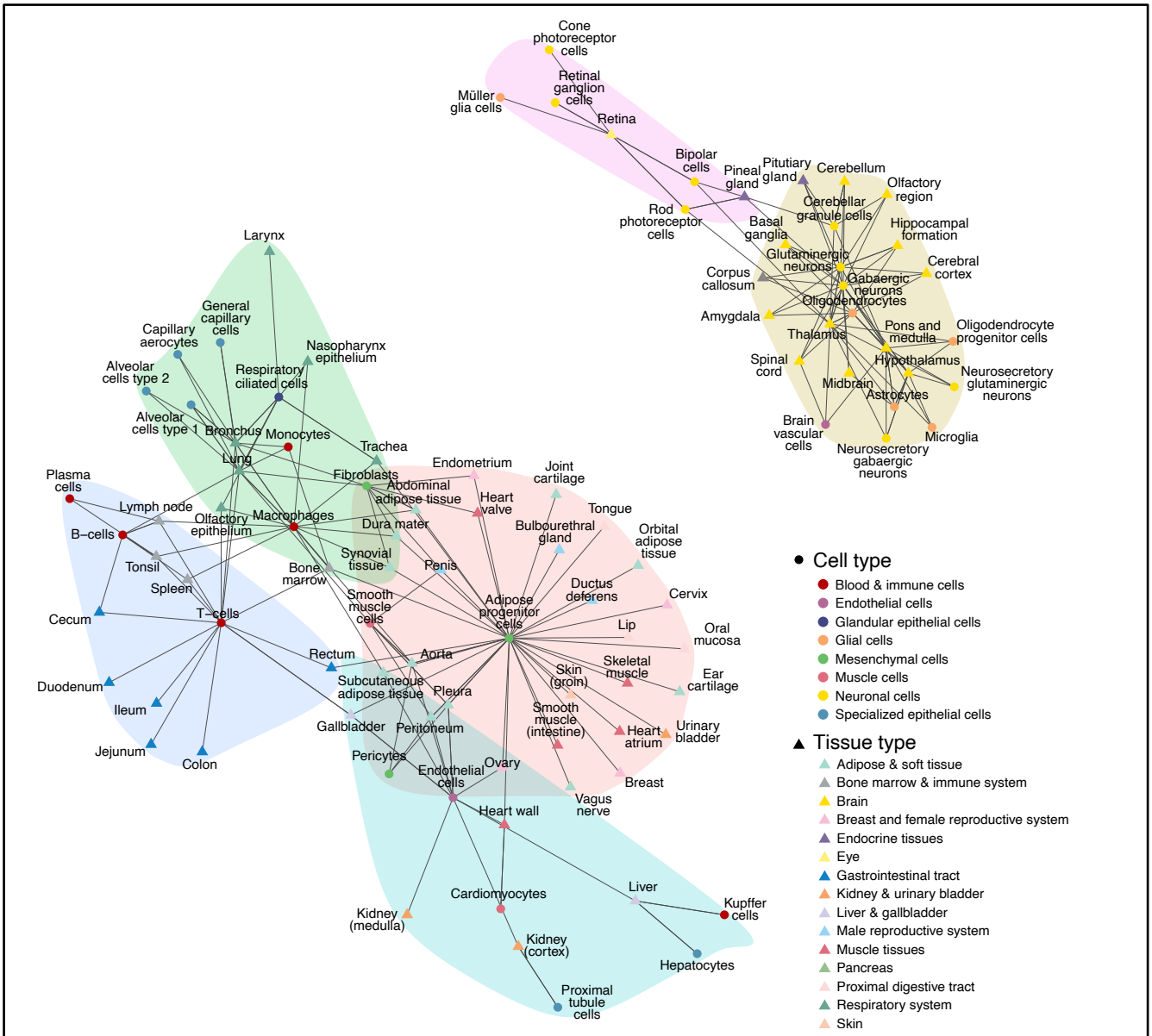


Figure 7. Network build upon the top 3 highest pairwise pearson correlations between cells types and tissue types based on enriched genes. All available tissue types of the bulk RNA atlas were included for this comparison.

high Tau score genes than the bulk dataset. This argues that annotating the specificity of genes based on their expression in single cells provides more information towards their specificity.

By comparing the number of matching elevated genes between the bulk and single cell data, one can compute a hypergeometric test and, in this way, assess the similarities between cell types and tissue types. (**Figure 6B**) This test serves as a control of the data to confirm both correct annotation and sampling of the single cell data. We can observe that mostly cell types can be traced back to their tissue of origin. A few cell types however do seem to have a significant overlap of enriched genes between with other tissues, such as the brain vascular cells with adipose tissue, ependymal cells with the lung, or various kidney cells with the liver tissue and liver cells with the kidney tissue. I investigated some of these relationships further through the atlas to corroborate them to be originating from genes involved in shared molecular processes of two tissues. (**Figure 9A**)

A similar test can be performed specific for brain tissues only. (**Figure S7**). The results of this test can hint again back to the population of neurons found in each brain region. We see Glutaminergic

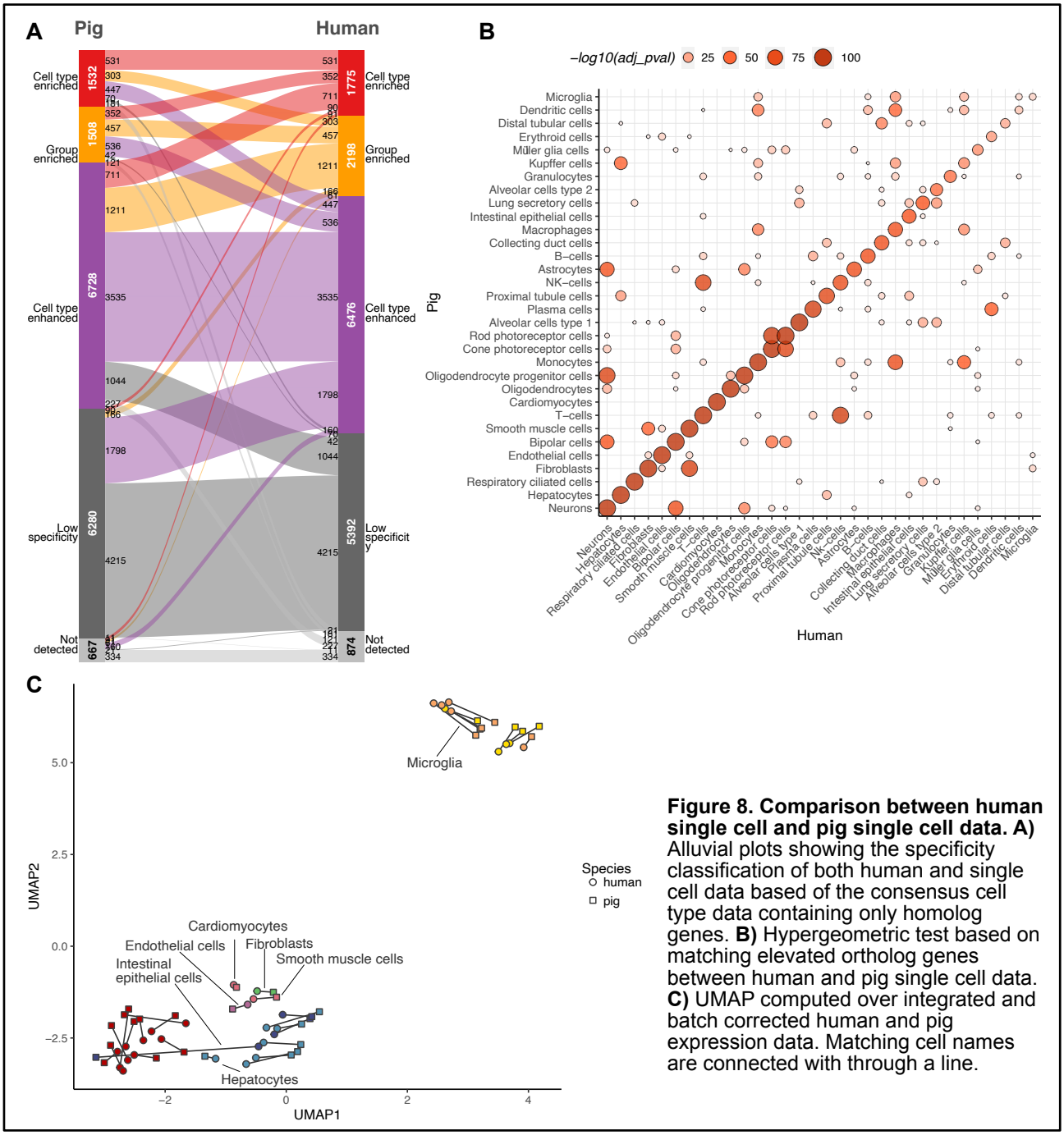


Figure 8. Comparison between human single cell and pig single cell data. A) Alluvial plots showing the specificity classification of both human and single cell data based of the consensus cell type data containing only homolog genes. **B)** Hypergeometric test based on matching elevated ortholog genes between human and pig single cell data. **C)** UMAP computed over integrated and batch corrected human and pig expression data. Matching cell names are connected with through a line.

and Gabarergic neurons to have significantly shared elevated genes between the amygdala, basal ganglia, cerebral cortex, hippocampal formation, and the olfactory region, however not in the cerebellum, corpus collosum, hypothalamus, midbrain or pons and medulla. Medium spiny neurons share elevated genes with the basal ganglia, the cerebellar cells with the cerebellum, and neurosecretory neurons with the hypothalamus and pons and medulla, the pituitary gland, and basal ganglia. The pineal gland and retina tissues seem to have also similar cell makeup due to, by matching elevated genes with the same cell types.

Finally, to assess how the complete bulk dataset compared to the cell type dataset of the pig, I calculated the pairwise Pearson correlations between cell type and tissue types and visualized this through a network plot as seen in **Figure 7**. Similarly, as in the hypergeometric test, we can see that cells correlate the most to tissues, in which they are present. For example, immune cells correlate highly with the lymph node and spleen, proximal tubule cells correlate with the cortex

correlate highly with the lymph node and spleen, proximal tubule cells correlate with the cortex region of the kidney, or hepatocytes and Kupffer cells correlate with the liver. Also, immune cells correlate highly with from tissues having mucosal surfaces like the respiratory or digestive tract. Adipose progenitor cells locate themselves in the center of connective tissues like the skin, muscle tissue, breasts, and soft tissues, aside from adipose tissues.

Comparison to human single cell data

Due to difference in cells detected and annotation, to compare the two datasets, I had to create a new consensus nomenclator between both tissue types. (**Figure S8**). I based on these names, I created a new comparison dataset for both pig and human single cell data, containing only ortholog genes (N = 16,715). With this in place, similarly as in the previous section, I classified all protein coding genes based on their specificity. (**Figure 8A**). We see that the number of genes assigned to each category are comparably similar.

With the annotation in place, to compare each consensus cell type more into detail I performed a hypergeometric test based on shared elevated genes. (**Figure 8B**). This test, like the hypergeometric test, one can control for correct labelling and highlight difference between the cell types. We see that every matching cell type share significantly overlap in matching enriched genes, as expected. Additionally, cell types undergoing similar processes share enriched genes as well. For example, microglia share overlap with immune cells like macrophages, Kupffer cells, and monocytes.

To compare the whole expression profile of both human and pig data, I visualized them in a UMAP plot. (**Figure 8C**) In this plot we can observe that mostly matching cell types do cluster together. The only outlier is the intestinal epithelial cells. The pig intestinal epithelial cells cluster together with blood and immune cells instead.

Discussion

Identification of cell types

The pig single cell atlas is built upon the single cell transcriptomic libraries of 24 samples from 11 tissues and 9 brain regions. The tissue sections in the dataset (**Figure 1A**) originate predominantly from functional organs, except for the adipose tissues, which are connective tissue, as well as the PBMC. The sections covered, are valuable for studies on cardiovascular disease, diabetes, metabolism, infectious and airborne diseases, cystic fibrosis, a variety of cancers, upon others.

The organ sections of the intestine, kidney, liver, and lung would be expected to be predominantly composed out of epithelial tissues featuring their organ specific parenchymal or specialized epithelial cells. For the tissue sections in the atlas, this holds true for the kidney, liver, and lung, however not for the intestine (**Figures 3, S1, S5**).

Consistent with a previous human kidney atlas published,[26] the kidney was predominantly populated by proximal tubule cells. Other cell types such as podocytes, loop of Henle cells, distal tubule cells, collecting duct cells are equally detected in kidney atlases. Missing in this atlas are

however glomerular parietal cells, which due to their usually low cell population and the low cluster definition in the pig sample might have not been able to form an independent cluster. To note, renal interstitial cells, are a mixture of mesenchymal cells in the kidney's stroma.

Similarly, for the cell annotation of the liver, cell types are consistent with what was identified in a human liver atlas.[27] The tissue is predominantly composed from hepatocytes. As reported in the human liver atlas, there is a predominant presence of Kupffer cells, T cells and other immune cells. Two populations of Kupffer cells were detected, potentially inflammatory and non-inflammatory Kupffer cells. Unfortunately, cholangiocytes were not detected, although due to observed low expression of EPCAM and ONECUT1 in the hepatocyte clusters of the liver snRNA-seq sample, that the clustering for them was not successful.

The scRNA-seq sample of the lung included an unexpected high number of immune cells, which would hint towards a flawed digestion process after sectioning. Only 98 non-immune cells (labelled as lung epithelial cells) from 10,906 cells were identified. The resulting cell population of snRNA-seq lung sample on the other hand was successful and is comparable to a previous human lung cell atlas[28] with regards to the cell types identified. As expected, alveoli epithelial cells alveolar type 1 and type 2 (AT1 and AT2) were identified, as well as respiratory ciliated cells of the airway epithelia. The cell type labelled as respiratory secretory cells include airway mucus secreting epithelial cells such as club cells or goblet cells that did not cluster separately. As for the human,[28] two main capillary cell types were identified: general capillary cells and the newly discovered[28] capillary aerocytes.

Like the lung scRNA-seq sample, the intestine sample seems to have failed the digestion process. Only 227 cells from 11,767 intestinal cells are non-immune. They were labelled as intestinal epithelial cells and endothelial cells. Unfortunately, due to the low number of cells, no higher resolution was obtainable. Additionally, 54 enteric glial and enteroendocrine cells were detected, however they were in a single cluster and could not be told apart. The cluster data was thus not included in the final expression dataset.

The heart is the only tissue in the atlas' sample set, that is expected to be predominantly muscle tissue. This was the case; the heart single cell library was composed predominantly from cardiomyocytes. The atrial cardiomyocytes marker MYL4 was not expressed in cardiomycyotes and marker NPPA only at low levels (7.4 nTPM), suggesting that these are only ventricular cardiomycyotes and that the heart atrium was not included in the tissue section. Aside from this, the cell population is like the one described in human heart single cell atlas[29], albeit the heart library includes a relatively low number of cells (N = 3,323).

Connective tissue is only represented by the two adipose tissue samples: visceral adipose (adipose-V) and subcutaneous adipose (adipose-S). Both samples are scRNA-seq samples, which consequently yielded no adipocyte to be detected. Adipocytes store fat, which is more buoyant than water and is thus escapes single cell sample preparation. Adipocytes are thus better detectable using snRNA-seq, however I had no access to a pig single nuclei adipose sample. Aside from missing adipocytes, both adipose tissues are consistent to human adipose single cell studies.[30] As in the human adipose atlas, visceral adipose contains mesothelial cells and a higher diversity in immune cells in comparison to subcutaneous adipose. Also as described in the human

adipose cell atlas, adipocyte progenitors, fibroblasts and mesenchymal stem cells are very closely related, and are difficult to tell apart due to their similarity.[30] They are here labeled as adipose progenitors.

The spleen is expected to be mostly populated out of a variety of immune cells, as it is part of the lymphatic system responsible for blood filtration. As expected, it includes a wide variety of immune cells. However, there were a few endothelial cells and muscle cells detected as well in the snRNAseq sample. The single nuclei sample identified additionally endothelial cells and smooth muscle cells which could originated from blood vessels in the section.

PBMC is known to be populated by lymphocytes (T-cells, B-cells, NK-cells), monocytes, and dendritic cells. This is consistent with what I have identified in the single cell library. Additionally, very few megakaryocytes and platelets were detected, which are not supposed to be found in the PBMC. However, as they are a minor component of the sample, it might be due to minor contamination during the isolation.

The 10 brain samples cover together a significant portion of the brain. (**Figure 1**) The atlas includes data from all four lobes of the cerebral cortex, the cerebellum, and the hypothalamus in both single cell and single nuclei. These together with the single cell data from the basal ganglia, thalamus, hippocampus, and pons cover a large amount of the brain. Included are in addition three of the four circumventrical organs: subformical organ, vascular organ of lamina terminalis (VOLT) and the area postrema. These are brain regions that are in contact with the blood and enable exchange between the central nervous system and the circulating blood and thus have endocrine function.[31]

The cell annotation for nerve tissues has been a technical and conceptual challenge in the field.[32] Annotation can vary based on neuron morphology or type of neurotransmitter produced. Neurons in retina single cell libraries however have been annotated mostly based on their morphology. I also took this approached and identified: rod cells, cone cells, ganglion cells and bipolar cells. Other retina atlases identify additionally amacrine cells and horizontal cells,[33,34] which were not identified in the pig libraries. These cells may have clustered however inside the ganglion cells clusters, since these clusters are expressing genes like TFAP2B, C1QL2 or GAD1, which are known markers for amacrine or horizontal cells in the retina.[33,34] Overall, the retina samples yielded poor cluster definitions, (**Figure S2**) which may have hindered the proper identification of these cells.

For the brain tissues, I decided to take a mixed approach. I aimed to define clusters of neurons by morphology if this was possible through transcriptomic data. Otherwise, I would define them primarily on if they produced predominantly GABA or glutamine as neurotransmitter as GABAergic or glutaminergic. If they did not express GABA or glutamine, then they would be defined by any other neurotransmitter they expressed, as cholinergic, serotinergic or dopaminergic. Additionally, due to the presence of the circumventrical organs, I labelled hormone secreting neurons as neurosecretory. By morphology, I identified cerebellar granule, purkinje and stellate cells, and medium spiny neurons.

Labeling neurons by neurotransmitter however proved to be challenging. The data showed often poor cluster definition, which is why re-clustering of the data was performed, to obtain the desired neuron cluster resolution. Additionally, it is normal for neurons to express simultaneously both

GABA and glutamine, or combinations of other neurotransmitters. I settled on labeling the neuron based on the highest expressed neurotransmitter, however this approach is not entirely reproducible and thus not ideal. Better suited annotation formats could be adopted from the single cell brain atlases such as the mouse brain atlas by Zeisel, et al.[35] Here, neuron clusters are given rather a code symbol as a label and annotated the (multiple) neurotransmitters as tags. Morphological information can be added as well. This approach might be however too complex and detailed for a multi-tissue single cell atlas.

To conclude on the cell type identification section, cell types in every organ were labelled and compared based on the available human single cell atlases of every organ. For most organs, the labelling was consistent, only the intestine sample, showed technical difficulties during tissue digestion. Recurrent problems in some samples however were poor cluster definition and low cell count number in some transcript libraries. These sometimes hindered on identifying some expected cell types in the samples. The method in annotating neurons could be still improved to be more reproducible.

A multi-organ cell type transcriptome map of the pig

Overall, the pseudo-bulk transcriptome data on 66 cell types did yield comprehensible cell type expression profiles. (**Figure 2**) The cell type clustering in **Figure 2C**, demonstrates mostly comprehensible cell type relationships, as annotated in the figure. Additionally, single cell data presents itself as a valuable complement to the bulk data, since genes are more specific to cell types than to tissues, (**Figure 6A, C**) which can aid methods like gene function annotation through gene clustering or generally give a more detailed insight of shared molecular processes between cell types. **Figure 6B** shows that single cell is indeed consistent with what was observed in the bulk tissue pig atlas.

Figure 6B highlights additionally interesting relationships between them. Kidney and Liver cells share enriched genes, due to their known shared task in detoxification. They coincide in the production of organic cation transport proteins, such as SLC22A1 (**Figure 9A**). Additionally, mucus secreting cells of the lung share enriched genes with the intestine; or ependymal cells in the brain carry cilia like the respiratory ciliated cells in the lung. **Figure S7** highlights some relationships of brain cells with brain regions. Medium spiny neurons are highly populated in the basal ganglia. The corpus collosum is predominantly populated of oligodendrocytes and not neurons; and the pinacocytes in the pineal gland are photoreceptors cells expressing melatonin similar to cells in the retina.[36]

Figure 8B suggests a correct labelling of cell types, since matching pig and human cells indeed show significant amount of matching elevated genes. However, intestinal epithelial cells clustering with myeloid cells in **Figure 2C** and **Figure 8C** highlights some issue of the dataset. Aside from the flawed intestine tissue digestion discussed in the previous section, it seems that there is significant ambient RNA present in the transcript data, which caused them to cluster together. Ambient RNA is understood as RNA residing outside of the cells, which can then be captured in solution together with the cell or nuclei in a droplet and can be labelled and amplified with the rest of the intracellular or intranuclear transcripts.[37] There could be similarly to the intestine sample, a high degree of ambient RNA in other samples. Suspects to this are for example also kidney and retina, both of which equally showed low cluster resolution and their cell types of clusters together in **Figure 2**.

Figure 8B gives further insight to what cells may have high ambient RNA contamination. For example, the expression data of pig astrocytes, overlaps highly with neurons. In human astrocytes, this is not the case since the astrocytes do not overlap as much with neurons. Similar can be said with hepatocytes and Kupffer cells. There is no overlap of human Kupffer cells with pig hepatocytes, however there is large overlap of pig Kupffer cells with human hepatocytes. This is also observed for pig plasma cells with erythroid cells, and pig intestinal epithelial cells with T cells.

A further issue with the dataset, is that despite the similarities and consistency across methods, the single cell atlas is not able to give a whole-body cell transcript picture of the pig. As mentioned, the library is primarily composed out of functional organs, and disregards connective tissue and muscle tissue. Adipose is the unique connective tissue in the dataset and this is potentially why adipose progenitor cells correlate the highest with multiple tissues (**Figure 7**). Cells rather expected to be there, include fibroblasts, smooth muscle cells or endothelial cells, but in the single cell dataset they were usually detected as part of vascular tissue and not connective tissue. The single cell dataset thus has rather a uncomplete picture of those cell types that are predominant connective tissues.

A navigable framework for single cell atlas exploration

Aside from going through the processing and labeling of data to generate a single cell, I had the objective of creating a navigable framework for the atlas by annotating all protein coding genes with cell type specificity categories and a Tau score, similarly as done previously by the Human Protein Atlas.[9,16] This framework should allow for efficient explorations of the genes in the atlas, while maintaining a wholistic view of the gene expression across cell types.

The practicality of this approach can be exemplified in the network plot in **Figure 5.** Through this framework, it is possible to swiftly find cell type markers through searching for cell type enriched genes. Similarly, it can enable to find shared enriched genes between different cell types, which would hint towards shared molecular processes. The other way around is also true. If a researcher is looking to validate involvement of a gene in a certain process across different cell types, this can be easily done by these categories, independent from the absolute nTPM expression values. This allows for streamlining of any kind of research involving pig as a model system. Multiple kinds of research areas can profit consequently benefit from this by facilitating hypothesis generation. From for example research on determining entry points of zoonotic diseases (**Figure 9B**) to determining novel disease markers to researching into creating gene function network in cells (**Figure 9C**).

Eventually, a visualization of gene expression of a gene for this single cell atlas could be presented as in **Figure 9**. In a single plot, one is able to acquire the information of the expression of a gene in a cell type but also specifically in what tissue this is expressed and if there are any changes in expression of this gene across tissues in the same cell type.

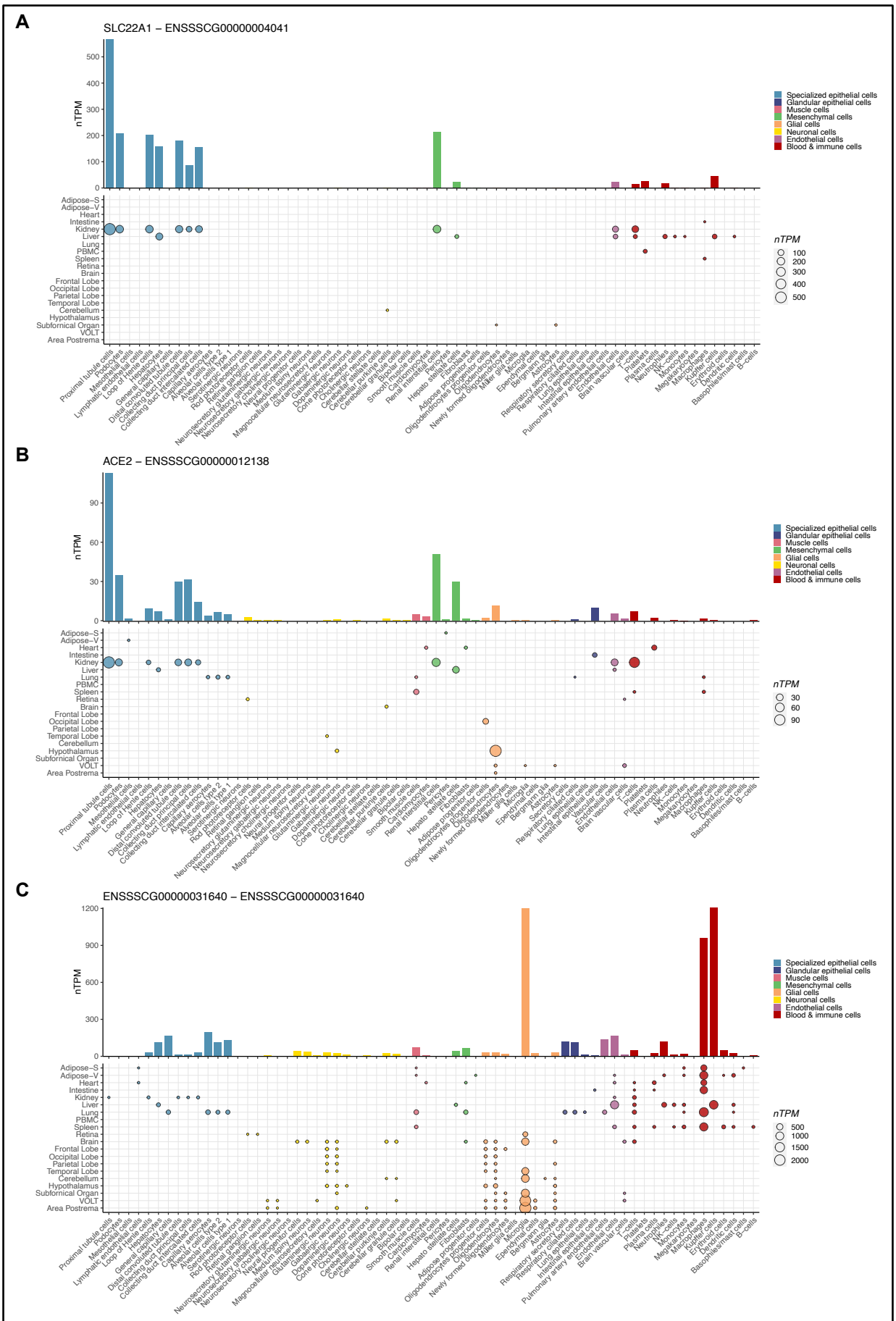


Figure 9. Examples of use cases and presentation of the pig single cell transcriptomic atlas. A) Genes expressed in shared tissues. SLC22A1, a cation transport protein, expressed in tissues of the lung and the liver. **B)** Mapping of viral entry points of zoonotic disease, by the example of the ACE2 receptor, entry point of the SARS-CoV-2 virus. **C)** Discovery of cell type specific genes exemplified by a group enriched gene in microglia, macrophages and Kupffer cells.

Future work

Throughout the discussion, I mentioned some aspects which could be improved upon in future work. In summary, there is some ambient RNA contaminating at least a portion of the cell types identified. Ambient RNA and low sample cell count may yield consequently poor cluster resolution and miss the identification of important cell types in a tissue. Furthermore, the tissues available for the atlas, were biased towards being functional organs, which in turn does not give complete picture of the cell types ubiquitous to connective and muscle tissue, such as fibroblasts, endothelial cells and smooth muscle cells.

As a short-term goal for this atlas, I would suggest for future work for this atlas, to run tools which remove ambient RNA contamination of the samples such a SoupX[37] or DecontX[38] and assess how this changes the overall atlas. It is a lengthy process since the whole labelling procedure would have to be performed again, but it might be a necessary step. This may help to establish higher cluster definition. I would suggest at that point also adjusting the framework of neuron labeling, to include the labelling of multiple neurotransmitters. This might involve creating a separate atlas for the pig brain only. This approach would otherwise yield considerably more cell types for the brain then for other tissues.

Additionally, a long-term goal, would be to try to counteract the bias for functional organs and complete the picture particularly for smooth muscle cells, fibroblasts, and endothelial cells. There diseases involving failure of muscle or connective tissues studied in pig, that would equally profit form this data, such as research in muscular dystrophy or osteosarcomas.[17] Similarly, due to pigs playing a central role growing organs for xenotransplatnation[17] researchers would profit, to have a baseline global cell type transcript reference of the pig. Similarly, the atlas would profit, if more than only one organism would be used per section to reduce bias.

Finally, as mentioned, this atlas will complement the bulk tissue atlas of the pig available at www.rnaatlas.org. In the future, this atlas will contain the reference transcriptomes at bulk and single cell level of various mammalian species, establishing ultimately the Mammalian RNA Atlas.

Ethical reflection

For this master thesis I report the process involved in generating a single cell atlas of the pig. The researchers that produced the data did report approval from the responsible ethical committees and have followed their applicable national and institutional animal use and welfare guidelines.

Additionally, a motivation of creating an easily comprehensible and navigable single cell atlas of the pig is to reduce the number of animals used in science. Having a reliable database can aid hypothesis generation or refute early hypothesis. This way, one may avoid unnecessary use of pigs and allow for a more effective use of them, with a better backed up hypothesis.

The research also features that at a cellular level, there are little differences between pigs and humans. Tissues and organs of both organisms are composed of the same cell types formed of equal building blocks. Pigs have complex social and cognitive capacities.[39] As mammals, both pigs and humans share the same biological necessities for well-being. By highlighting the similarities between pig and human at a cellular level, might argument for improved standards of animal use and animal welfare.

Acknowledgements

I would like to acknowledge following people for their support during the degree project:

My supervisor Linn Fagerberg, for allowing me to collaborate with the Human Protein Atlas for my degree project and providing me with insightful advice towards how to steer the whole project. I appreciate her confidence in me, allowing for very independent work however and her being swiftly supportive for any issue, no matter how last minute this was.

Mengnan Shi, for introducing me to the single cell pipeline analysis and helping me out when I had any single cell data processing related question.

Max Karlsson, for showing how to use R to make great plots and introducing me to the bulk data analysis pipeline.

Jan Mulder and Evelina Sjöstedt, for their helpful discussions on cell type labeling and reasoning on my results.

References

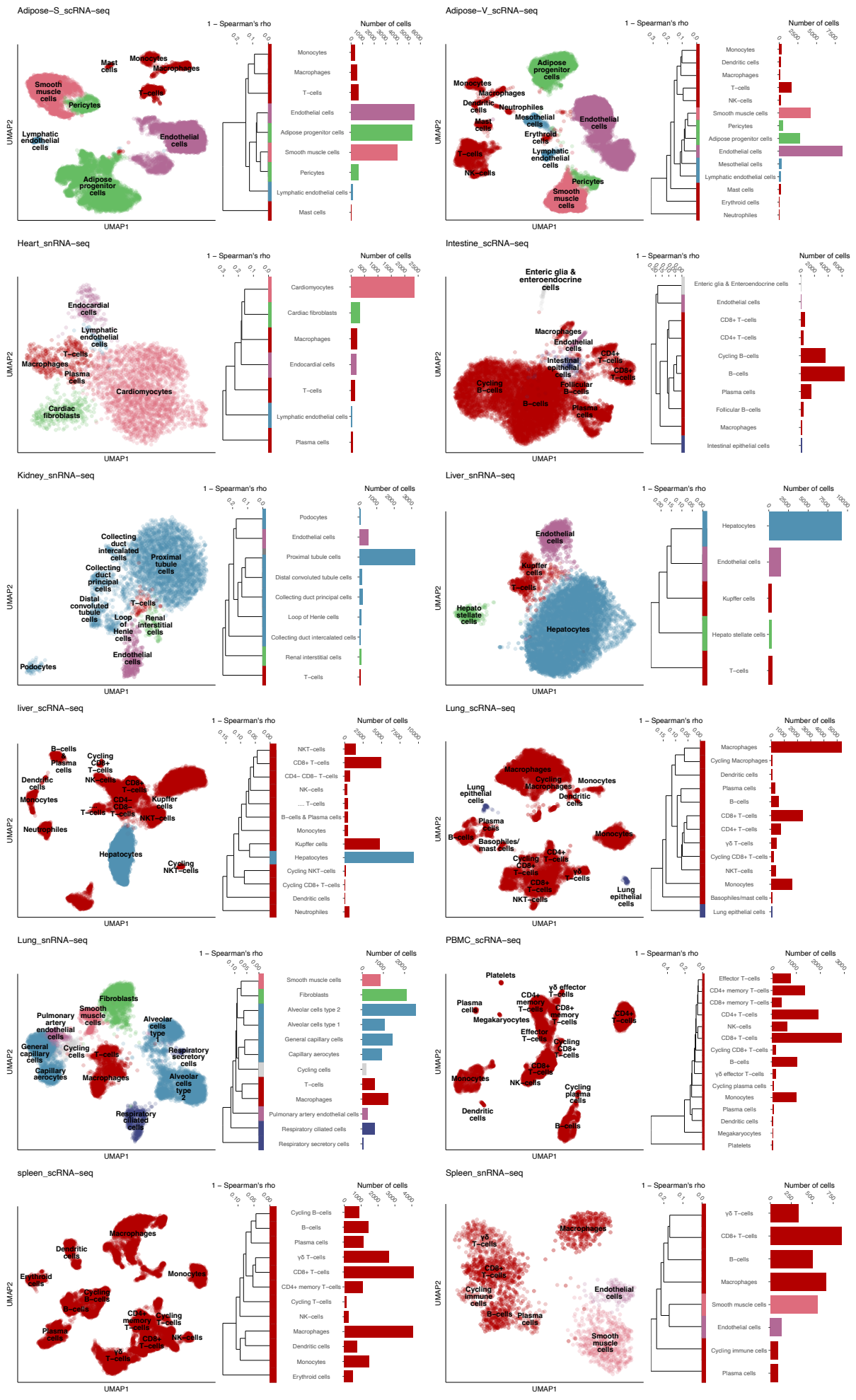
1. Ramón y Cajal, S. *Histologie du système nerveux de l'homme et des vertébrés*. (1909).
2. Hershey, B. A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *Journal of General Physiology* **36**, 39–56 (1952).
3. Crick, F. On Protein Synthesis. *The Symposia of the Society for Experimental Biology* 138–163 (1958).
4. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (1979)* **291**, 1304–1351 (2001).
5. International Human Genome Sequencing Consortium *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, (2001).
6. Brenner, S. *Sydney Brenner - Nobel Lecture: Nature's Gift to Science*. (2002).
7. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).
8. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160–1167 (2011).
9. Karlsson, M. *et al. A single-cell type transcriptomics map of human tissues*. *Sci. Adv* vol. 7 www.proteinatlas.org (2021).
10. Jones, R. C. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (1979)* **376**, (2022).
11. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics* Preprint at https://doi.org/10.1038/s41576-023-00580-2 (2023).
12. Adil, A., Kumar, V., Jan, A. T. & Asger, M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Frontiers in Neuroscience* vol. 15 Preprint at https://doi.org/10.3389/fnins.2021.591122 (2021).
13. Dai, X. & Shen, L. Advances and Trends in Omics Technology Development. *Frontiers in Medicine* vol. 9 Preprint at https://doi.org/10.3389/fmed.2022.911861 (2022).
14. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nat Med* **28**, 2486–2496 (2022).
15. Chen, D. *et al.* Single cell atlas for 11 non-model mammals, reptiles and birds. *Nat Commun* **12**, (2021).
16. Karlsson, M. *et al.* Genome-wide annotation of protein-coding genes in pig. *BMC Biol* **20**, (2022).
17. Lunney, J. K. *et al. Importance of the pig as a human biomedical model*. *Sci. Transl. Med* vol. 13 https://www.science.org (2021).
18. Gutierrez, K., Dicks, N., Glanzner, W. G., Agellon, L. B. & Bordignon, V. Efficacy of the porcine species in biomedical research. *Front Genet* **6**, (2015).
19. Wang, F. *et al.* Endothelial cell heterogeneity and microglia regulons revealed by a pig cell landscape at single-cell level. *Nat Commun* **13**, (2022).
20. Zhu, J. *et al.* Single-cell atlas of domestic pig cerebral cortex and hypothalamus. *Sci Bull (Beijing)* **66**, 1448–1461 (2021).
21. Zhang, L. *et al.* A high-resolution cell atlas of the domestic pig lung and an online platform for exploring lung single-cell data. *Journal of Genetics and Genomics* **48**, 411–425 (2021).
22. Warr, A. *et al.* An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* **9**, (2020).
23. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).
24. Robinson, M. D. & Oshlack, A. *A scaling normalization method for differential expression analysis of RNA-seq data*. http://genomebiology.com/2010/11/3/R25 (2010).

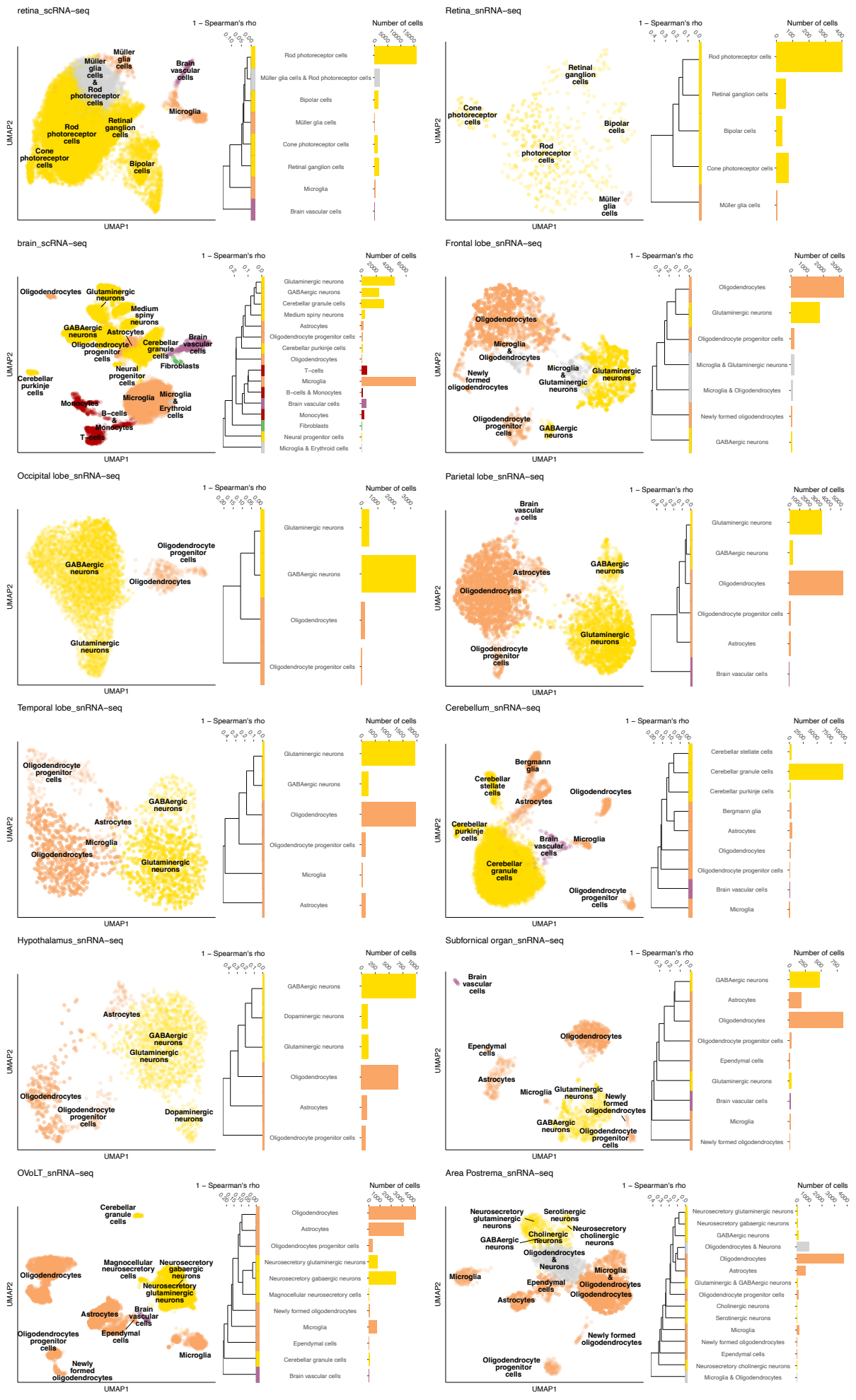
25. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
26. Liao, J. *et al.* Single-cell RNA sequencing of human kidney. *Sci Data* **7**, (2020).
27. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, (2018).
28. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
29. Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466–472 (2020).
30. Norreen-Thorsen, M. *et al.* A human adipose tissue cell-type transcriptome atlas. *Cell Rep* **40**, (2022).
31. Ganong, W. F. Circumventricular organs: Definition and role in the regulation of endocrine and autonomic function. in *Clinical and Experimental Pharmacology and Physiology* vol. 27 422–427 (2000).
32. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: Challenges, opportunities and the path forward. *Nature Reviews Neuroscience* vol. 18 530–546 Preprint at https://doi.org/10.1038/nrn.2017.85 (2017).
33. Liang, Q. *et al.* A multi-omics atlas of the human retina at single-cell resolution. *Cell Genomics* 100298 (2023) doi:10.1016/j.xgen.2023.100298.
34. Menon, M. *et al.* Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nat Commun* **10**, (2019).
35. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).
36. Pandi-Perumal, S. R. *et al.* Melatonin: Nature's most versatile biological signal? *FEBS Journal* vol. 273 2813–2838 Preprint at https://doi.org/10.1111/j.1742-4658.2006.05322.x (2006).
37. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, (2020).
38. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol* **21**, (2020).
39. Warr, A. *et al.* An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* **9**, (2020).

Supplementary Figures

Figures bellow.

Figure S1 & S2. Data presentation of the 24 samples individually. For every sample, a UMAP plot, a dendrogram and a cell count barplot was computed. The UMAP plot, as calculated by the scanpy preprocessing pipeline and labelled with the cluster names. The dendrogram shows the relationship between the assigned cell types inside the sample. The bar plot pictures the number of cells detected per cell type inside a sample.


Adipose−S_scRNA−seq
1 − Spearman's rho
Number of cells
Mast cells
Monocytes
Macrophages
Smooth muscle cells
T−cells
Pericytes
Endothelial cells
Lymphatic endothelial cells
Adipose progenitor cells
Monocytes
Macrophages
T−cells
Endothelial cells
Adipose progenitor cells
Smooth muscle cells
Pericytes
Lymphatic endothelial cells
Mast cells
UMAP1
UMAP2
Adipose−V_scRNA−seq
1 − Spearman's rho
Number of cells
Adipose progenitor cells
Monocytes
Macrophages
Dendritic cells
Neutrophiles
Mesothelial cells
Mast cells
Erythroid cells
Endothelial cells
T−cells
NK−cells
Lymphatic endothelial cells
Pericytes
Smooth muscle cells
Monocytes
Dendritic cells
Macrophages
T−cells
NK−cells
Smooth muscle cells
Pericytes
Adipose progenitor cells
Endothelial cells
Mesothelial cells
Lymphatic endothelial cells
Mast cells
Erythroid cells
Neutrophiles
UMAP1
UMAP2
Heart_snRNA−seq
1 − Spearman's rho
Number of cells
Endocardial cells
Lymphatic endothelial cells
T−cells
Macrophages
Plasma cells
Cardiomyocytes
Cardiac fibroblasts
Cardiomyocytes
Cardiac fibroblasts
Macrophages
Endocardial cells
T−cells
Lymphatic endothelial cells
Plasma cells
UMAP1
UMAP2
Intestine_scRNA−seq
1 − Spearman's rho
Number of cells
Enteric glia & enteroendocrine cells
Macrophages
Endothelial cells
CD4+ T−cells
Intestinal epithelial cells
CD8+ T−cells
Follicular B−cells
Cycling B−cells
B−cells
Plasma cells
Enteric glia & Enteroendocrine cells
Endothelial cells
CD8+ T−cells
CD4+ T−cells
Cycling B−cells
B−cells
Plasma cells
Follicular B−cells
Macrophages
Intestinal epithelial cells
UMAP1
UMAP2
Kidney_snRNA−seq
1 − Spearman's rho
Number of cells
Collecting duct intercalated cells
Proximal tubule cells
Collecting duct principal cells
Distal convoluted tubule cells
T−cells
Loop of Henle cells
Renal interstitial cells
Endothelial cells
Podocytes
Podocytes
Endothelial cells
Proximal tubule cells
Distal convoluted tubule cells
Collecting duct principal cells
Loop of Henle cells
Collecting duct intercalated cells
Renal interstitial cells
T−cells
UMAP1
UMAP2
Liver_snRNA−seq
1 − Spearman's rho
Number of cells
Endothelial cells
Kupffer cells
T−cells
Hepato stellate cells
Hepatocytes
Hepatocytes
Endothelial cells
Kupffer cells
Hepato stellate cells
T−cells
UMAP1
UMAP2
liver_scRNA−seq
1 − Spearman's rho
Number of cells
B−cells & Plasma cells
Cycling CD8+ T−cells
NK−cells
CD8+ T−cells
Dendritic cells
Monocytes
CD4− CD8− T−cells
Kupffer cells
Neutrophiles
NKT−cells
Hepatocytes
Cycling NKT−cells
NKT−cells
CD4+ T−cells
Kupffer cells
Hepatocytes
Cycling NKT−cells
Cycling CD8+ T−cells
Dendritic cells
Neutrophiles
UMAP1
UMAP2
Lung_scRNA−seq
1 − Spearman's rho
Number of cells
Macrophages
Cycling Macrophages
Monocytes
Lung epithelial cells
Dendritic cells
Plasma cells
B−cells
Basophiles/mast cells
Cycling CD8+ T−cells
CD4+ T−cells
γδ T−cells
NKT−cells
Macrophages
Cycling Macrophages
Dendritic cells
Plasma cells
B−cells
CD8+ T−cells
CD4+ T−cells
γδ T−cells
Cycling CD8+ T−cells
NKT−cells
Monocytes
Basophiles/mast cells
Lung epithelial cells
UMAP1
UMAP2
Lung_snRNA−seq
1 − Spearman's rho
Number of cells
Fibroblasts
Smooth muscle cells
Pulmonary artery endothelial cells
Alveolar cells type 1
General capillary cells
Cycling cells
T−cells
Respiratory secretory cells
Capillary aerocytes
Macrophages
Alveolar cells type 2
Respiratory ciliated cells
Smooth muscle cells
Fibroblasts
Alveolar cells type 2
Alveolar cells type 1
General capillary cells
Capillary aerocytes
Cycling cells
T−cells
Macrophages
Pulmonary artery endothelial cells
Respiratory ciliated cells
Respiratory secretory cells
UMAP1
UMAP2
PBMC_scRNA−seq
1 − Spearman's rho
Number of cells
Platelets
γδ effector T−cells
Plasma cells
CD4+ memory T−cells
CD8+ memory T−cells
Megakaryocytes
Effector T−cells
CD4+ T−cells
Cycling CD8+ T−cells
CD8+ T−cells
Monocytes
NK−cells
Cycling plasma cells
Dendritic cells
B−cells
Effector T−cells
CD4+ memory T−cells
CD8+ memory T−cells
CD4+ T−cells
NK−cells
CD8+ T−cells
Cycling CD8+ T−cells
B−cells
γδ effector T−cells
Cycling plasma cells
Monocytes
Plasma cells
Dendritic cells
Megakaryocytes
Platelets
UMAP1
UMAP2
spleen_scRNA−seq
1 − Spearman's rho
Number of cells
Macrophages
Dendritic cells
Erythroid cells
Monocytes
Cycling B−cells
B−cells
Plasma cells
CD4+ memory T−cells
Cycling T−cells
NK−cells
γδ T−cells
CD8+ T−cells
Cycling B−cells
B−cells
Plasma cells
γδ T−cells
CD8+ T−cells
CD4+ memory T−cells
Cycling T−cells
NK−cells
Macrophages
Dendritic cells
Monocytes
Erythroid cells
UMAP1
UMAP2
Spleen_snRNA−seq
1 − Spearman's rho
Number of cells
γδ T−cells
CD8+ T−cells
Cycling immune cells
B−cells
Plasma cells
Macrophages
Endothelial cells
Smooth muscle cells
γδ T−cells
CD8+ T−cells
B−cells
Macrophages
Smooth muscle cells
Endothelial cells
Cycling immune cells
Plasma cells
UMAP1
UMAP2


retina_scRNA−seq
1 − Spearman's rho
Number of cells
Müller glia cells
Müller glia cells & Rod photoreceptor cells
Brain vascular cells
Microglia
Retinal ganglion cells
Rod photoreceptor cells
Cone photoreceptor cells
Bipolar cells
Müller glia cells
Cone photoreceptor cells
Retinal ganglion cells
Microglia
Brain vascular cells
UMAP1
UMAP2
Retina_snRNA−seq
Retinal ganglion cells
Cone photoreceptor cells
Bipolar cells
Rod photoreceptor cells
Müller glia cells
brain_scRNA−seq
Oligodendrocytes
Glutaminergic neurons
Medium spiny neurons
GABAergic neurons
Astrocytes
Oligodendrocyte progenitor cells
Cerebellar granule cells
Brain vascular cells
Fibroblasts
Neural progenitor cells
Cerebellar purkinje cells
Microglia
Microglia & Erythroid cells
Monocytes
B−cells & Monocytes
T−cells
Frontal lobe_snRNA−seq
Oligodendrocytes
Microglia & Oligodendrocytes
Newly formed oligodendrocytes
Microglia & Glutaminergic neurons
Oligodendrocyte progenitor cells
GABAergic neurons
Occipital lobe_snRNA−seq
Oligodendrocyte progenitor cells
Parietal lobe_snRNA−seq
Brain vascular cells
Temporal lobe_snRNA−seq
Cerebellum_snRNA−seq
Cerebellar stellate cells
Bergmann glia
Cerebellar granule cells
Cerebellar purkinje cells
Hypothalamus_snRNA−seq
Dopaminergic neurons
Subfornical organ_snRNA−seq
Ependymal cells
OVoLT_snRNA−seq
Magnocellular neurosecretory cells
Neurosecretory gabaergic neurons
Neurosecretory glutaminergic neurons
Area Postrema_snRNA−seq
Neurosecretory glutaminergic neurons
Serotinergic neurons
Cholinergic neurons
Neurosecretory cholinergic neurons
Oligodendrocytes & Neurons

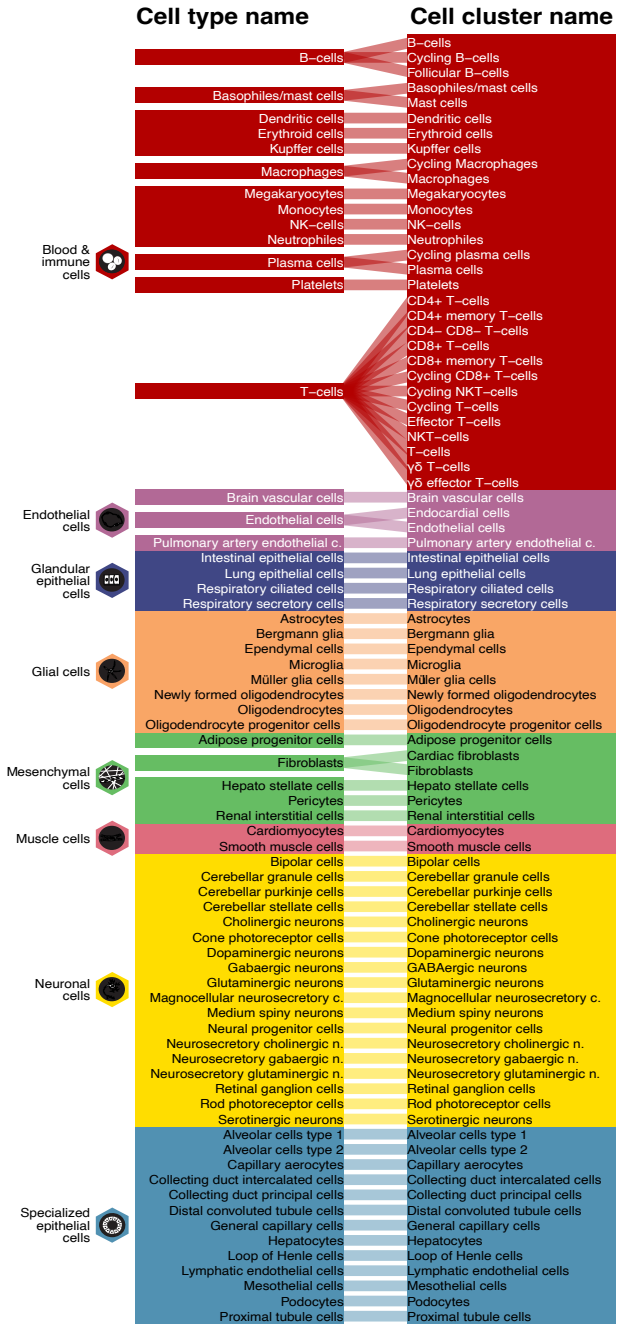


Figure S3. Renaming strategy for cell cluster names into cell type names.

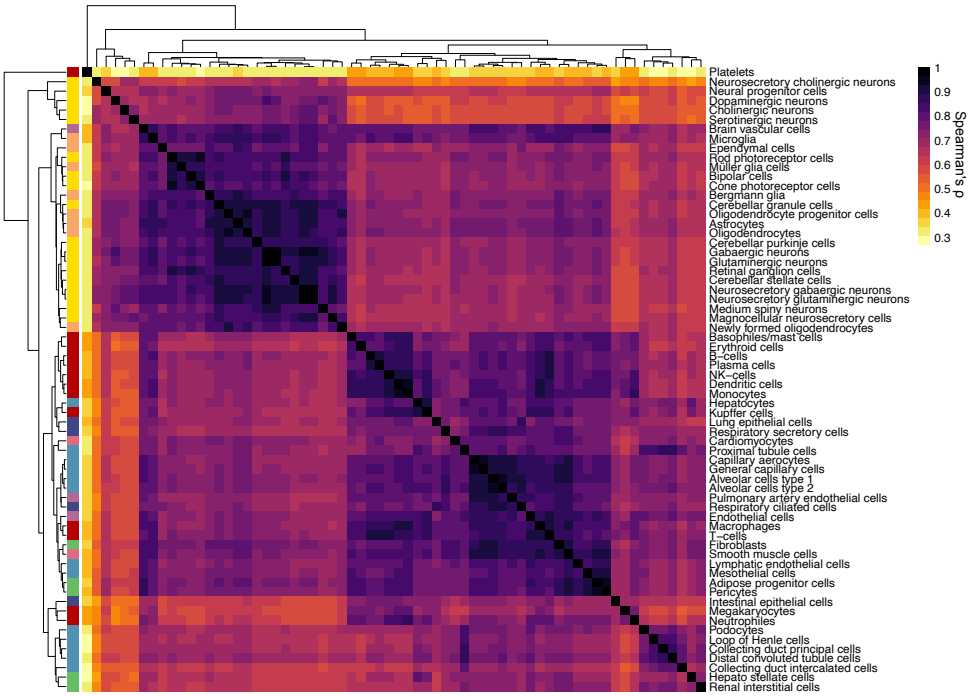


Figure S4. Clustered heatmap showing the pairwise spearman correlation between each cell type. Clustering calculated with complete linkage method.

Figure S5. Relative cell count identified for each cell type by tissue. A total of 66 cell types were identified across all 24 libraries or 20 tissues. Relative cell count was calculated as the quotient of the counts per cell type in a tissue and the total cell count of a tissue. For samples, which had both single cell and single nuclei libraries, the total cell count of a tissue was computed through first adding the total cell count in a sample together.

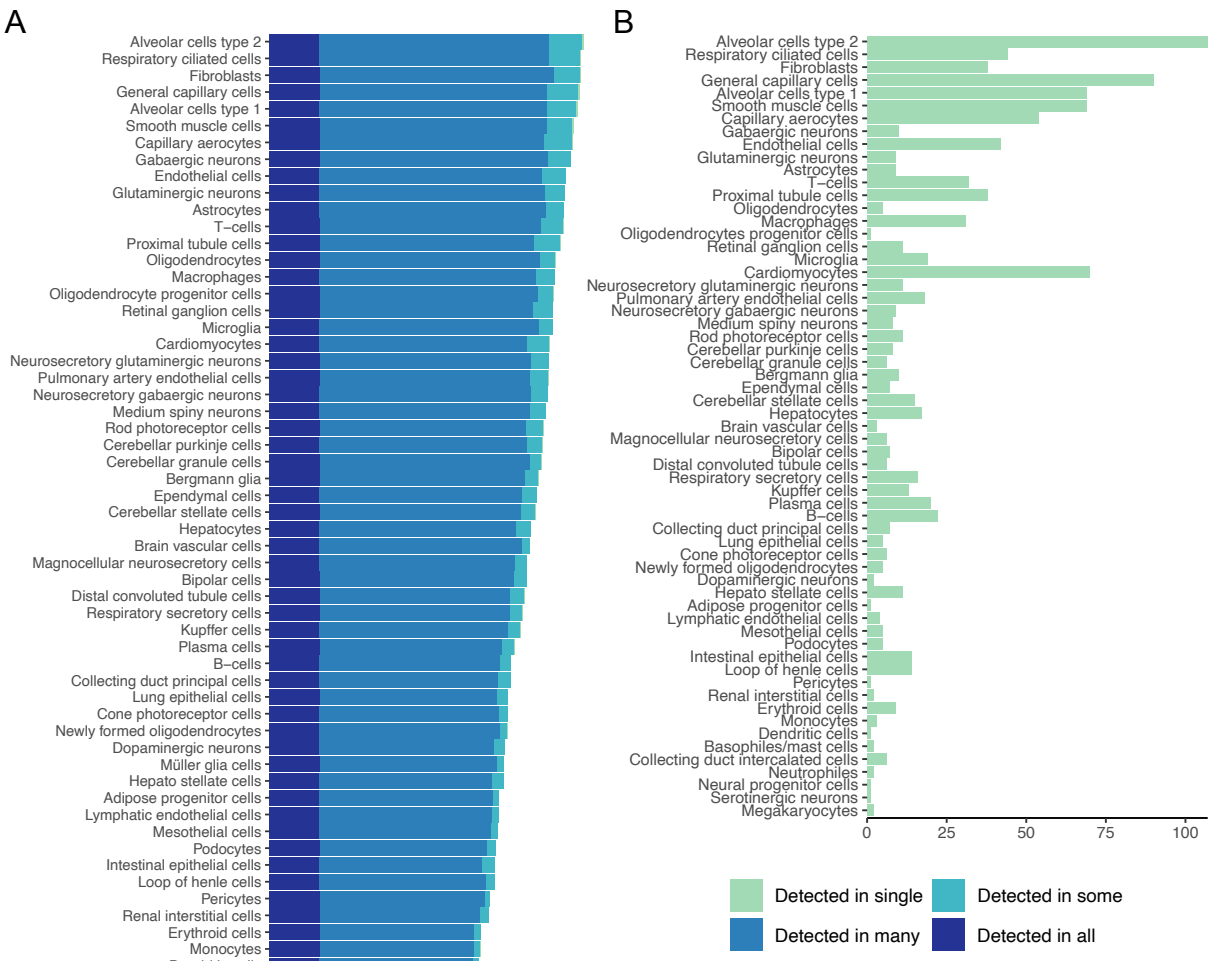


Figure S6. Gene distribution annotation annotation per cell type. A) For every cell type, the amount of genes annotated to each of the four distribution categories. B) Number of genes annotated as detected in single for for each cell type.

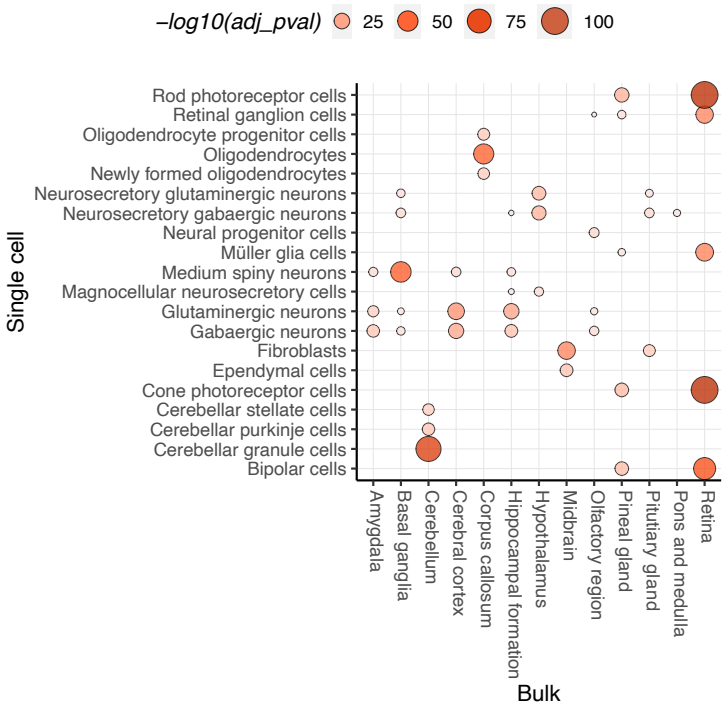


Figure S7. Hypergeometric test based on matching elevated genes between cell types detected in brain tissue and retina and bulk RNA tissue of regions located in the brain and the retina.

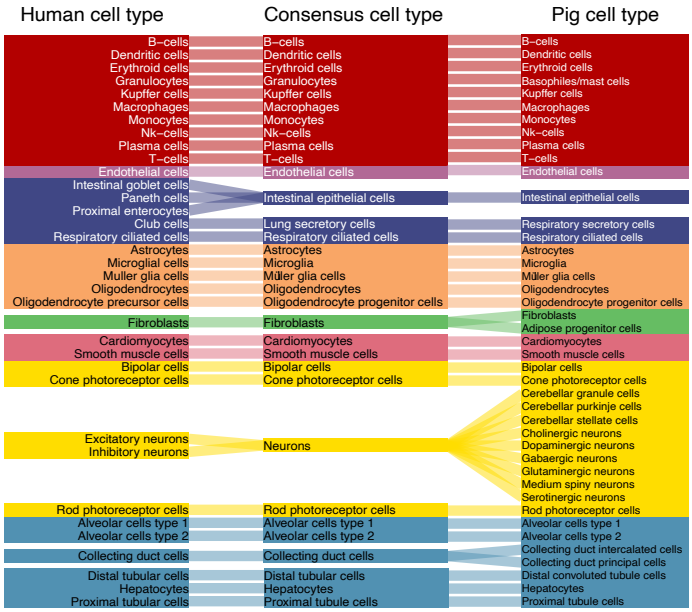


Figure S8. List of cell types pooled together in order to create a matching consensus cell type gene expression matrix used as a basis for the human to pig single cell type comparison. The consensus datasets were computed by taking the mean expression of every gene in a cell type assigned to matching consensus cell types. Only ortholog genes were included.