Project Report
CB2050

Rat RNA atlas: Bulk RNAseq data analysis, gene annotation and human cross species comparison

Supervised by
Linn Fagerberg

Submitted by
Emilio Skarwan
skarwan@kth.se

15th January, 2023
MSc Molecular Techniques in Life Science
KTH Royal Institute of Technology

Contents

Abstract

The rat is widely used model organisms used for scientific research applied in most areas of the molecular life sciences and pharmacology. However, there is still no accessible resource where to find baseline gene expression levels of the rat. In this project report I present my work done over the last 10 weeks on the data analysis for the development of the Rat RNA Atlas at Mathias Uhlén's lab under supervision of Linn Fagerberg. The RNA Atlas dataset is composed of transcriptomes of 361 samples from 100 distinct tissues across the whole body and it is set to become a database to explore and compare the gene expression profiles of protein coding genes of the rat in different tissues spanning the whole body. I present my workflow starting with filtering for protein coding genes and data normalization through trimmed means of M values (TMM) of transcript per million (TPM) values for tissue comparison. This is followed by quality control, which concluded in the reduction of the dataset to 352 samples. The project followed with annotation of all protein coding genes in terms of tissue specificity and distribution using discrete categories as well as a continuous Tau score for tissue specificity; and culminated with a cross species comparison of gene expression profiles between the rat and the human. With this work I conclude that the dataset with the adjustments presented here is of quality to become a RNA Atlas of the rat, and that the assigned annotation is essential for an efficient and effective usage of a gene expression database.

Introduction

It is well established at latest in the central dogma[1] that the protein gene expression pathway typically goes through the transcription of DNA in the cell nucleus to RNA, which is maturated and exported to the cytosol and translated into protein, which then carry out diverse functions in the cell. While the DNA is mostly identical across the whole organism, due to heterogeneous gene expression regulation the RNA content varies depending on the cell type and even then, has temporal variation depending on cell cycle state and random fluctuations.[2]

As an analogy, one could say that RNA expression profile of a tissue or cell is comparable to the short-term memory or RAM of a computer. As the data in the RAM is made up of data currently used by the computer at a given moment, so do the protein coding RNA transcripts give an insight to what genes are expressed in each tissue or cell. Thus, exploring cross-tissue RNA expression patterns of an organism's body is essential for the understanding of the processes happening in any given section and so further understanding their molecular similarities and differences between other sections. In this way giving insight to understanding the whole organism's biology, from a sum of smaller specialized sub-sections.

Throughout the years in the molecular life sciences and pharmacology the rat has been established as an important model system. The earliest reported scientific experiments on a rat go back to as early as 1856, with Philipeaux and a rat adrenaloctomy.[3] Its significance has only grown over time as the rat has become an essential model animal employed in basic molecular biology, biomedical and pharmacological research. The 2019 European Commission report on scientific animal use report that rats account to 9.4 % of all animals used for scientific purposes in the European Union and Norway, falling second place behind mice (52.5 %).[4]

However, in spite of their wide use and scientific relevance, an easily accessible database containing baseline (i.e. not differential expression) RNA expression profiles of multiple tissues spanning the whole body is yet to be established. The EMBL-EBI expression atlas, a repository centralizing over 3,000 Gene expression experiments of different organisms in one place, lists only three baseline RNA-seq experiments on the Rat.[5] The RatBodyMap being the largest one spanning only eleven organs, albeit through different developmental stages.[6]

Thus, having an accessible resource containing RNA-sequencing data of several tissues spanning the whole body of the rat would be a valuable tool with for future research. This motivated an expansion of the Human Protein Atlas to create a Mammalian RNA atlas.[7] The mammalian RNA atlas (rnaatlas.org) is currently solely composed of the Pig, but it will expand to more mammals, continuing by the rat.

I present consequently here my work contributing to the creation of a Rat tissue atlas, which involved data processing of bulk RNA sequencing data of 100 different tissues spanning the whole body of the rat, followed by analysis and gene annotation regarding tissue specificity and distribution across tissues and finalizing with a cross-species comparison of the tissue specific gene expression profiles between Human and Rat. Similar to what Karlsson et al. presented with the Pig Atlas[7], and following the presentation and format initially presented in the Human Protein atlas database.[8]

Methods

Sample Collection, library preparation and sequencing.

The sample collection and handling of the animals was not by me, which is why I cannot get into the details of it. However, tissue samples (N = 361) come from 8 *Rattus norvegicus*, four males and four females. Similarly, RNA extraction, library preparation and sequencing were not part of my project.

Processing of raw data and data normalization

The raw fastq output was processed through Kallisto and mapped to *Rattus norvegicus* Ensemble build version 103. Kallisto,[9] an RNA algorithm based on read pseudoalignment, generated the transcript count data as well as a TPM (transcript per million) normalisation. The TPM normalized

data was filtered to only contain protein coding transcripts. Following this, a new per million normalisation is calculated as pTPM (protein TPM) values. The pTPM for each gene was pooled together by addition and the expression data went further through TMM normalization (Trimmed means of M values).[10] This returns then our final normalized TPM values or nTPM.

Generation of expression dataset at different hierarchies
The transcriptomic dataset of the Rat is organized in different hierarchies: tissue types are assigned to region tissues, which are assigned to grouped tissues that are organized under organ systems. For each gene in each tissue type, the mean expression over all tissue replicates is taken. The region dataset is generated by taking the maximum value of each gene inside each region. This hierarchy is however only relevant for brain tissues. The tissues are also grouped into grouped tissues, the maximum expression is taken for each gene inside a grouped tissue, as in the region tissue dataset.

Gene distribution and specificity categorization and tau score calculation
Every gene is classified regarding tissue specificity and distribution based on their nTPM values. For tissue specificity there are five categories established: tissue enriched, group enriched, tissue enhanced, low tissue specificity and not detected. Tissue enriched genes are genes which in a single tissue have a 4-fold nTPM expression value compared to any other tissue. Group enriched genes are genes expressed in 2-5 tissues with an nTPM expression level higher than a fourth of the maximum expression and an average nTPM value of at least 4-fold higher than the rest of the tissues. Tissue enhanced genes are genes that have at least 4-fold higher nTPM expression than the average expression in a group of tissues. Tissue enriched, group enriched and tissue enhanced genes are collectively defined as tissue elevated genes. Low tissues specificity is assigned to genes that fail to be assigned the previous specificity categories. Not detected genes, are genes having an nTPM value under 1.

Similarly, there are five gene distribution categories: Detected in all are genes showing a nTPM ≥ 1 in all tissues. Detected in many are genes showing a nTPM ≥ 1 in at least 31% of tissues, but not in all. Detected in some are genes with an nTPM of ≥ 1 in more than 1 tissue but less than 31% of tissues. Detected in single are genes with an nTPM ≥ 1 for a single tissue. Not detected are genes with nTPM < 1 in all tissues, however annotated in the ensemble 103 genome.

Additionally for each gene a tau (τ) score or tissue specificity index was calculated using log10(nTPM + 1) transformed data. The tau score is calculated as defined by Ynai, et al. [11], where N is the number of tissues and x_i is the expression value relative to the highest expression of a gene.

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

Comparison of rat and human orthologs
For the cross-species comparison between Rat and Human transcriptome, the human transcriptomic data was downloaded from the Human Protein Atlas (HPA) available under “RNA consensus tissue gene data” accessible at https://www.proteinatlas.org/about/download. The dataset contains nTPM expression values of all protein coding genes from 54 tissues based on data of the HPA and GTEx based on Ensemble 103.

For comparison, the data was batch corrected using the removeBatchEffect function the Limma package on R. A dataset containing only ortholog genes (n = 16,157) of both organisms was filtered, based on a curated high confidence ortholog list between human and rat provided by Kalle von Feilitzen of the HPA based off ensemble 103.

Principal component analysis (PCA)
For the PCA plots, log10(nTPM+1) transformed values were taken and center scaled. Principal components were calculated using the pca function in the pcaMethods package in R.

Uniform Manifold Approximation and Projection (UMAP)
For UMAP plots, the principal components were calculated as described above. Later PCA components accounting to 80% of the variability were selected. On these selected PCs umap was performed using the umap function in the uwot R package. UMAP was calculated with 15 neighbors for 1000 epochs using the Euclidean distance metric.

Spearman correlation and spearman distance (Spearman ρ)
For the calculation of the spearman correlation, the function cor was used found inside the stats base package in R. For this function the method "spearman" was selected. Spearman distance is 1 – spearman correlation.

Clustered heatmaps
Clustered heatmaps were built using the pheatmap function available in the pheatmap package. Clustering method used was the Euclidean method.

Hypergeometric test
A hypergeometric test was performed to compare the tissue elevated genes in human with the elevated genes in rat. For this, both comparison datasets composed only of orthologous genes were to be categorized in terms of tissue specificity. Number of tissue elevated genes inside each tissue were then extracted. For the test each rat tissue was compared against each human tissue. The function phyper from the stats base package in R was used. For this q was defined as the number of orthologs elevated in both the human and the rat tissue; m was the number of enriched orthologs of the tissue with the lowest number between the two tissues being compared; n was the total number of orthologs – m; k was the total number of orthologs elevated in either tissue or in both. This test was repeated until all rat tissues were tested against all human tissue. FDR (false discovery rate) was then calculated from the p value by using the Benjamini Hochberg p adjustment method available as "BH" under p.adjust function of the stats base package in R.

Data analysis and visualization:
Data analysis and visualization was performed on R version 4.2.1 (2022-06-23) through RStudio 2022.07.2. Most visualizations were plotted using ggplot2 (v 3.4.0) and slightly edited on Affinity Designer 2 for aesthetics and labeling purposes. Other R packages used were: influential (v 2.2.6), geomtextpath (v 0.1.1) , ggsci (v 2.9), viridis (0.6.2), viridisLite (v 0.4.1) patchwork (v 1.1.2) ggraph (v 2.1.0), ggrepel (v 0.9.2), ggplotify (v 0.1.0), pheatmap (v 1.0.12), uwot (v 0.1.14.9000), Matrix (v 1.5-3), ggdendro (v 0.1.23), ggalluvial (v 0.12.3), forcats (v 0.5.2), stringr (v 1.5.0), dplyr (v 1.0.10) purr (v 0.3.5), readr (v 2.1.3), tidyr (v 1.2.1), tibble (v 3.1.8), tidyverse (v 1.3.2), biomaRt (v 2.52.0) pcaMethods (v 1.88.0), Biobase (v 2.56.0) and BiocGenerics (v 0.42.0). Code used to plot figures is available at https://github.com/emiliosk/RatRNAAtlas_CB2050/tree/main

Results

Overview of the Rat RNA Atlas dataset
The final transcriptomic dataset of the Rat RNA Atlas is composed of 352 samples, which together are assigned to 100 individual tissue types. The tissues are grouped into 92 region tissues, this hierarchy level however is only relevant for the brain tissues. The tissues are further grouped into 53 grouped tissues based on their function and development, as well as trying to keep a similar grouping as the Pig RNA atlas. Similarly, the 53 grouped tissues are then assigned one of 15 organ systems. **Figure 1A** shows the relationships between tissue types, grouped tissues and organ systems. **Figure S1** shows the region tissue to region relationship for the brain tissues.

The reads were mapped on the reference, only protein coding transcripts were selected, transcripts were pooled by gene and the normalized TPM (nTPM) values were calculated as described in the methods section. **Figure S2** shows the effect of TMM normalization on TPM data. After normalization, I generated three further expression datasets at different hierarchy levels from the baseline sample nTPM expression dataset. This results then in four expression datasets: nTPM by sample, nTPM by tissue, nTPM by region and nTPM by group.

Before further downstream analysis however I had to go through quality control (QC) to ensure right labeling and consistency of the dataset. The quality control plots used on the initial QC step can be seen under **additional file 1**. Through this plots identified nine aberrant samples which we had to remove ("adis_female.1", "adis_female.4", "adis_male.1", "adis_male.4", "care_male.1", "cor_female.3", "cor_male.3", "corpus.callosum_male.2", "entorhinal.cortex_female.2", "len_male.2"). The dataset got reduced from 361 to 352 samples. Also, there are 3 sample pairs that I concluded as having been switched during the previous steps and therefore had their names exchanged in the sample expression dataset. These were: “occipital.cortex_female.2” and “midbrain_female.2”; “midbrain_female.2” and “occipital.cortex_female.2”; “ute_female.4” and “end_female.1”; “vagi_female.4” and “cer_female.1”; “end_female.1” and “ute_female.4”; “cer_female.1” and “vagi_female.4”). These were investigated thoroughly through different means, such as using dendrograms based on spearman distance and heat maps showing expression of gene markers enriched for a certain tissue.

With the quality control step done and the changes implemented to the dataset I could continue with further analysis. Still, a second set of quality control plots were calculated with the final dataset for final validation (**additional file 2**). 94.9% of genes (20,657 genes out of 22,245) were detected (nTPM ≥ 1) at least in one tissue. For each tissue, the detection ranges from a minimum of 13,698 detected genes for lens to 15,794 detected genes for olfactory epithelium. (Figure S2)

To assess the similarities relationship between the whole tissues, I visualized the pairwise Pearson correlation between all tissues as a dendrogram built using Ward’s minimum variance clustering. **(Figure 1B).** It is visible how most tissues assigned to one organ system are clustered together. One can observe for example the brain tissues, muscle tissues, gastrointestinal tract tissues and connective, soft tissues and bone marrow & lymphoid tissues clustered together. Other tissues show to be clustered due to their proximity inside the body, such as the intestine smooth muscle clustered close to the gastrointestinal tract tissues, or the Tear gland tissue falling together with the Skin (eye) tissue, and other cluster together due to related functions or cellular composition. However, the location of cornea and the fallopian tube tissues are not immediately explicable due to this.

Additionally, to show the correlation between tissue groups I present a clustered heatmap showing the pairwise Pearson correlation (**Figure 1C**). Testis, Brain and Retina show to be the most distinct of all grouped tissues. A similar plot at the tissue level is provided in the supplemental (**Figure S3**) shows similarly Testis to be the most have the lowest correlation to all other samples, as well as all brain tissues except for dorsal root ganglia, to have very similar expression patterns with themselves but very distant to other tissues. This relationship is similarly visible in the circular dendrogram in **Figure 1B**.

Next, the expression profiles of the 352 samples were depicted through UMAP and PCA dimensionality reductions (**Figure 2**). These highlight again relationships between samples already observed in the previous plots. The brain samples show to show a very distinct expression profiles compared to the other samples. Interestingly in the UMAP (**Figure 2B**) one can recognize two brain clusters. One composed of the cerebral cortex, hippocampal formation, amygdala, basal ganglia, olfactory bulb and hypothalamus samples. While a second one composed of cerebellum, spinal chord, pons and medulla, midbrain, corpus callosum and thalamus samples. (**Figure 2C).** Aside from the Brain cluster, one can identify the Gastrointestinal tract cluster, which also includes the Smooth muscle (intestine) samples. In the Gastrointestinal tract one can also detect three sub-clusters, which are the large intestine, small intestine, and stomach samples respectively. The most distant cluster is the liver sample cluster.

In general, one can see in both the UMAP (**Figure 2B**) as well as the PCA (**Figure 2A**) plots that samples from the same organ for the most show similar expression profiles. This can be assessed more in detail by plot showing Pearson correlation of samples to their same organ systems, to different organs systems, as well as the samples correlation to their own assigned tissue as seen in **Figure 2D**. All samples correlate highly to other samples of their own tissue, and there is a lower sample to sample correlation between different organs than in the same organ. Additionally, a detailed plot to sample to tissue correlation for every tissue is also depicted. There is an

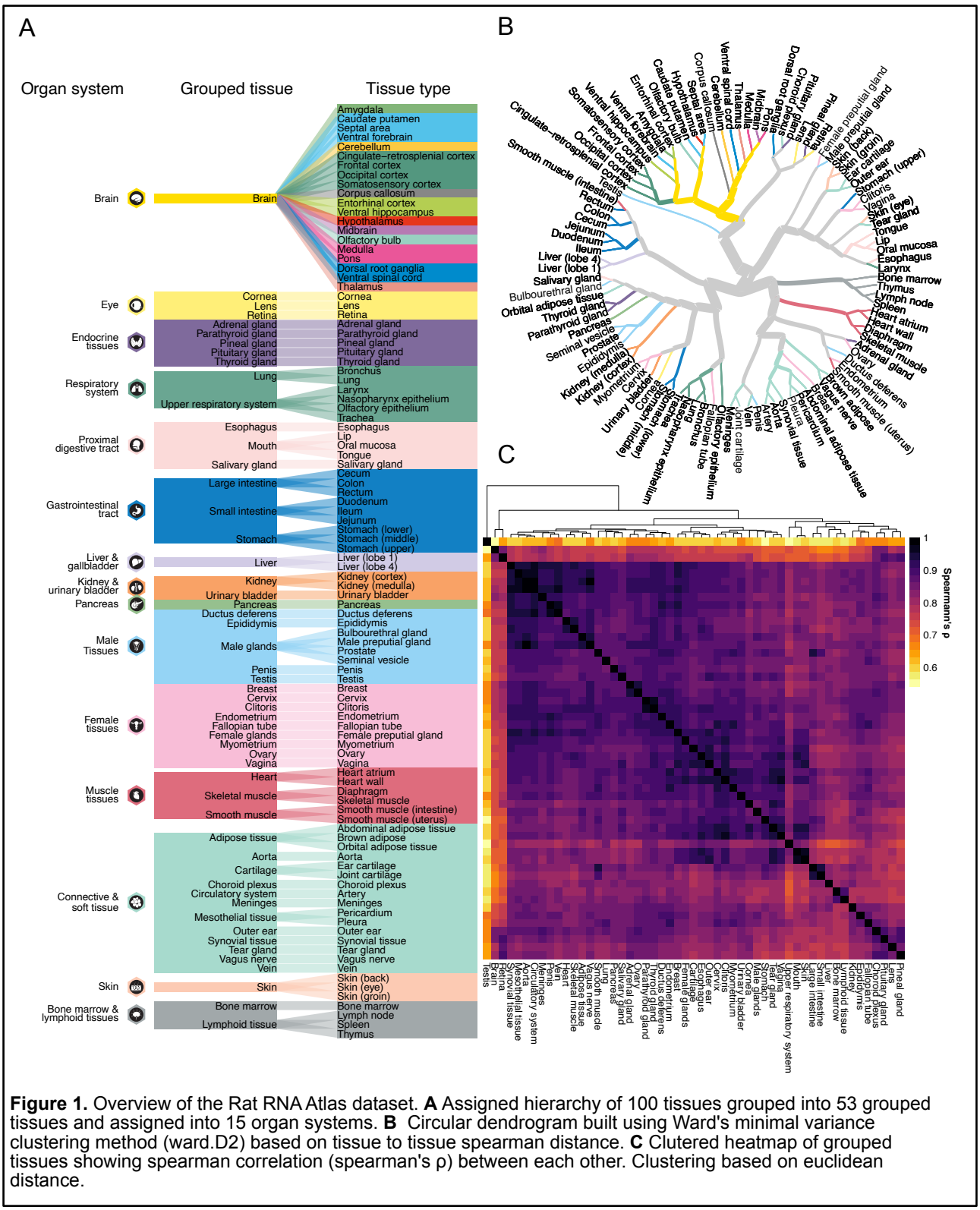

Figure 1. Overview of the Rat RNA Atlas dataset. **A** Assigned hierarchy of 100 tissues grouped into 53 grouped tissues and assigned into 15 organ systems. **B** Circular dendrogram built using Ward's minimal variance clustering method (ward.D2) based on tissue to tissue spearman distance. **C** Clutered heatmap of grouped tissues showing spearman correlation (spearman's ρ) between each other. Clustering based on euclidean distance.

average sample to assigned tissue correlation of 0.98 and ranges from 0.89, corresponding to a smooth muscle (intestine) sample to 0.99, a hippocampus sample.

Genome wide annotation of Rat protein coding genes

Similarly, as done before in the human protein atlas and in the pig atlas[7] we proceeded to annotate the 22,245 genes in terms of tissue distribution and tissue specificity. (**Figure 3** and **Figure 4**). The tissue distribution categories as described more in detail in the methods section are detected in single, detected in some, detected in many, detected in all and not detected. The tissue distribution categories shown in **Figure 3** were calculated at the tissue hierarchy level.

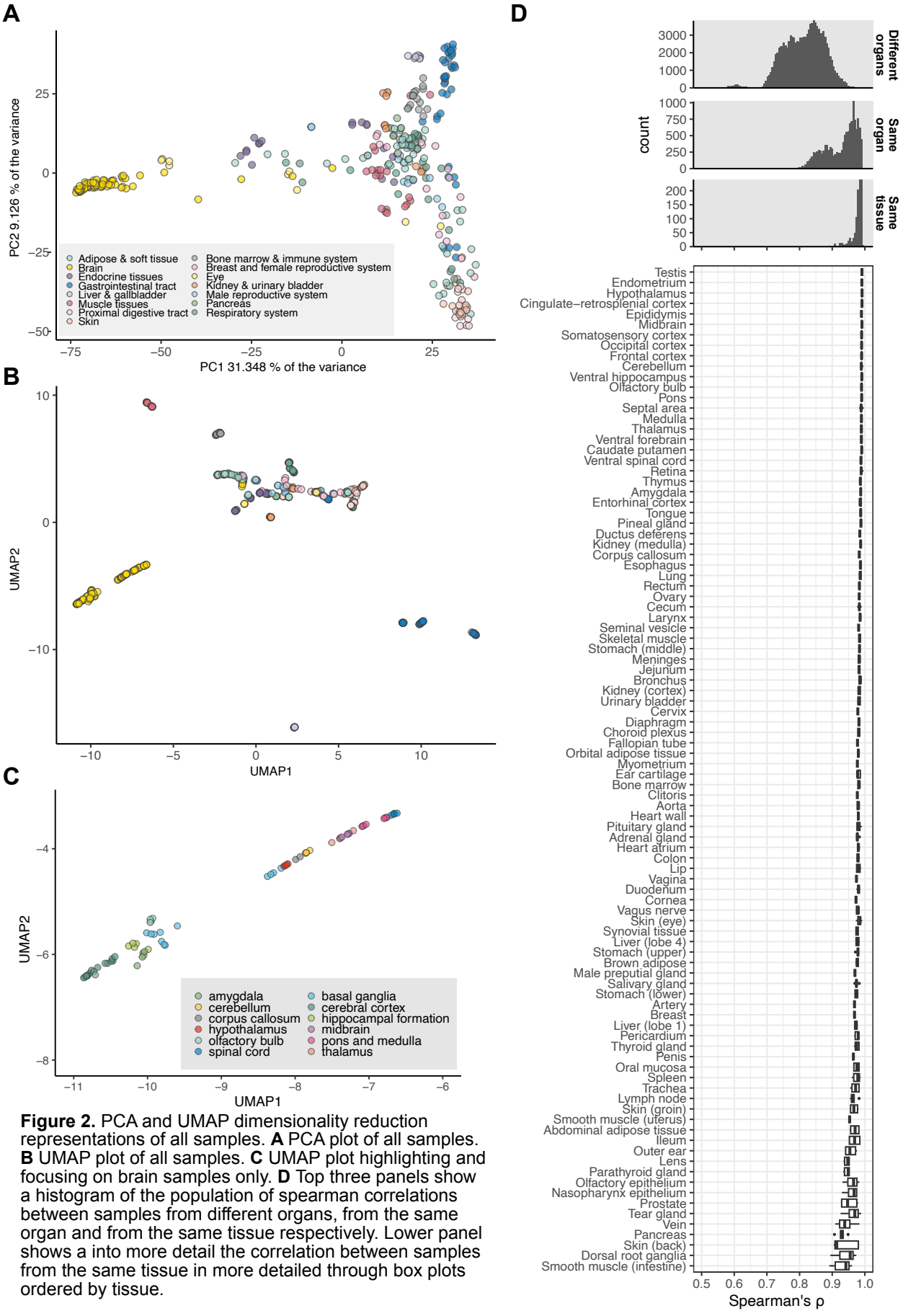

Figure 2. PCA and UMAP dimensionality reduction representations of all samples. **A** PCA plot of all samples. **B** UMAP plot of all samples. **C** UMAP plot highlighting and focusing on brain samples only. **D** Top three panels show a histogram of the population of spearman correlations between samples from different organs, from the same organ and from the same tissue respectively. Lower panel shows a into more detail the correlation between samples from the same tissue in more detailed through box plots ordered by tissue.

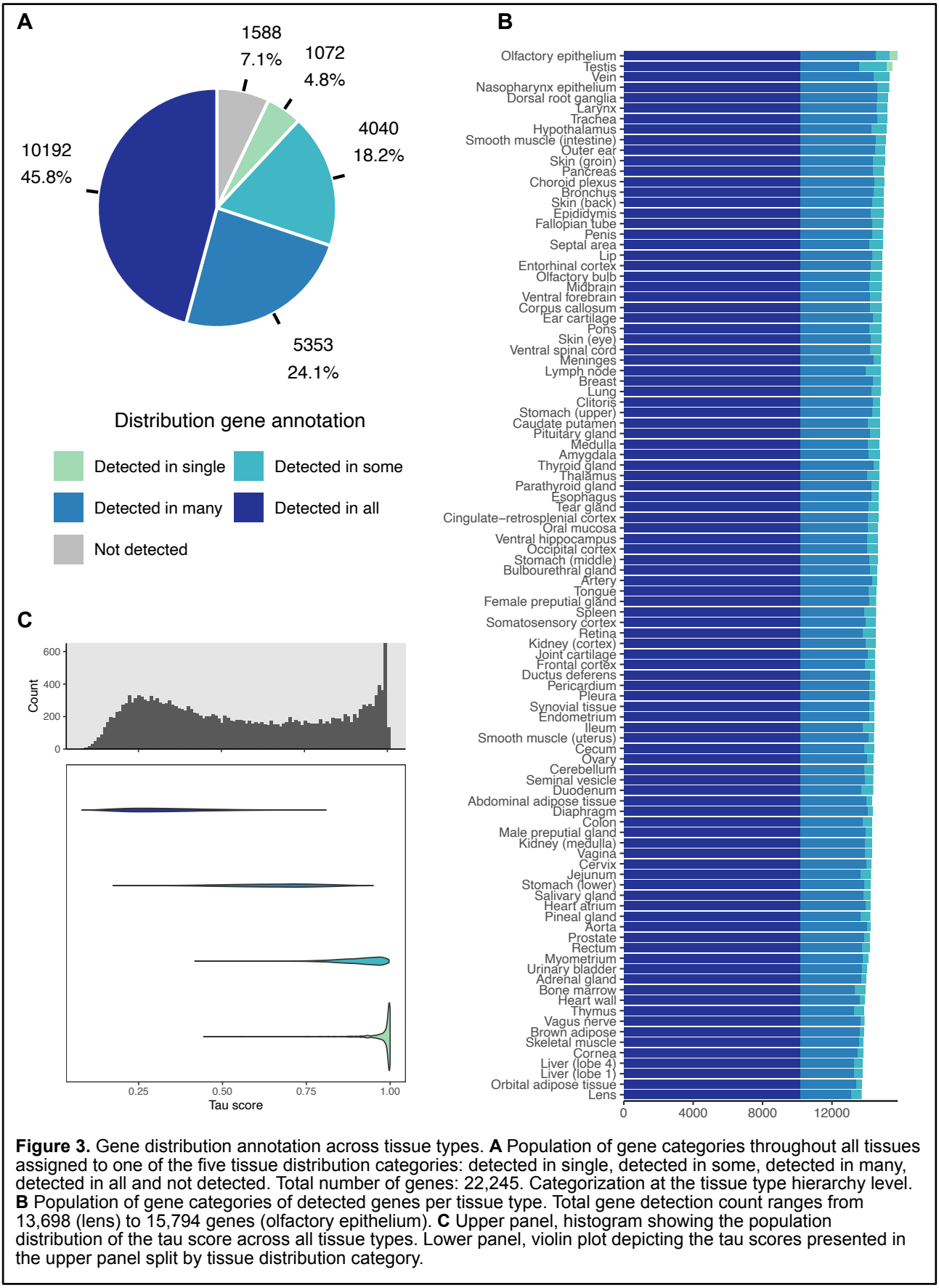

Figure 3. Gene distribution annotation across tissue types. **A** Population of gene categories throughout all tissues assigned to one of the five tissue distribution categories: detected in single, detected in some, detected in many, detected in all and not detected. Total number of genes: 22,245. Categorization at the tissue type hierarchy level. **B** Population of gene categories of detected genes per tissue type. Total gene detection count ranges from 13,698 (lens) to 15,794 genes (olfactory epithelium). **C** Upper panel, histogram showing the population distribution of the tau score across all tissue types. Lower panel, violin plot depicting the tau scores presented in the upper panel split by tissue distribution category.

Figure 3A shows that 20,657 out of 22,245 genes (92.9%) were detected. 46% of all genes (or 49% of all detected genes) were detected in all tissues. Only 4.8% or 1,072 genes were detected in a single tissue. Most of those concentrated on the olfactory epithelium and the testis which

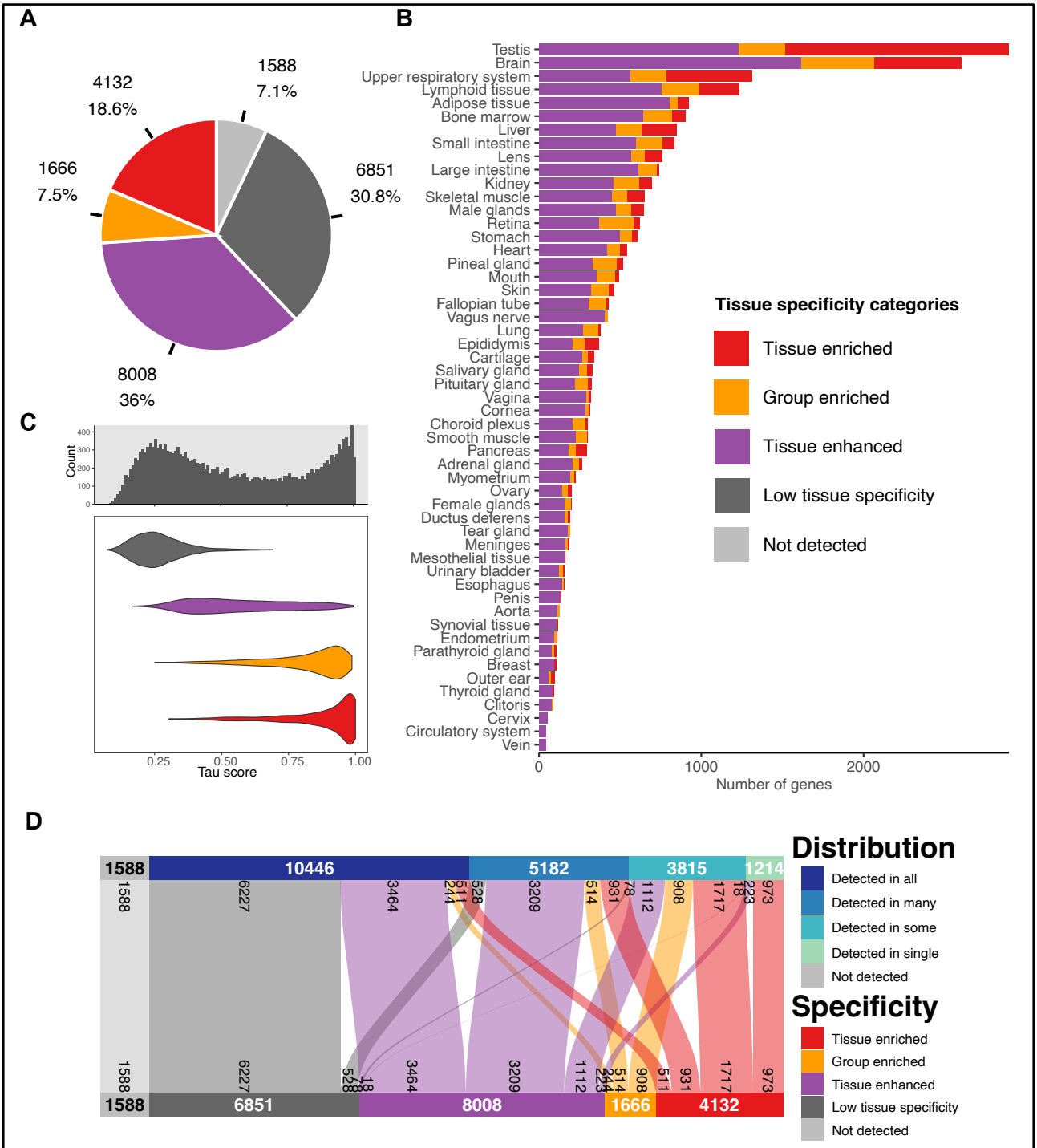

Figure 4. Tissue specificity annotation across grouped tissues. **A** Population of tissue specificity gene categories. Each gene is assigned one of five tissue specificity categories: tissue enriched, group enriched, tissue enhanced, low tissue specificity or not detected. Total number of genes: 22,245. Categorization at the grouped tissue hierarchy level. **B** Populations of tissue specificity categories showing only tissue elevated genes. Number of tissue elevated genes range from 40 (vein) to 2896 (testis) genes. **C** Upper panel, histogram showing the population distribution of the tau scores across all grouped tissues. Lower panel, violin plot depicting the tau scores presented in the upper panel split up by specificity distribution category. Tau score for non-detected genes (nTPM < 1) to NA. **D** Alluvial plot depicting the number of genes assigned to each tissue distribution and specificity category and the resulting overlap. Both distribution and specificity are here calculated at the grouped tissue level.

also show to have the highest number of detected genes (15,794 and 15472 respectively). (**Figure 3B**) The lowest number detected genes are detected for the lens tissue, although it is the 9th tissue with highest genes detected in a single tissue amounting to 12 genes. Only 71 of the 100 tissues show to have genes detected in that single tissue, but all 100 tissues have genes

categorized as detected in some. **Figure 3C** presents the population of all the gene's tau scores. As one would expect, detected in single tissue genes have the highest tau score, while detected in all have a lower one.

The specificity categories ordered in highest specificity to lowest specificity are tissue enriched, group enriched, tissue enhanced, low tissue specificity and not detected. Details on their definition are to be found in the methods section. Tissue enriched, group enriched, and tissue enhanced are also collectively referred to as tissue elevated. (**Figure 4)** Specificity categorization was done at the tissue group hierarchy. 13,806 genes show to be tissue enhanced (62.1%), while 6,851 genes show low tissue specificity (30.8%) (**Figure 4A**). **Figure 4B** shows the specificity categorization for genes in each grouped tissue. Testis (N = 2,896) and brain (N = 2,600) show to have the highest number of tissue elevated genes, while vein the lowest (N = 40). Overall, the results are very similar to what had been observed in the pig. **Figure 4C** depicts the relationship between tissue specificity categories and the tau score. As expected, we can observe a high tau score for highly tissue specialized genes, and a low score for low specific genes. The alluvial plot in **Figure 4D** shows the overlap between tissue distribution and tissue specificity of both categorization at the grouped tissue level.

Finally, to better depict the gene tissue specificity and relationship between tissues and tissue elevated genes, I present a network plot. (**Figure 5**) The network plot shows the numbers of either tissue enriched in red nodes or group enriched in orange nodes connected to their respective grouped tissues. Since tissue enriched is defined enriched in a single tissue, it is only connected to one grouped tissue node. However, group enriched genes can be enriched in up to five tissues per definition, and each tissue enriched node can be connected to up to four grouped tissues, which depicts the number of grouped enriched genes shared between grouped tissues. For example, most enriched genes are shared between bone marrow and lymphoid tissues (N = 102), which could be explained due to their high content on immune cells and immune cell maturation. Second highest number of grouped enriched genes (N = 49) are shared between the brain and the retina, followed by shared enriched genes (N =47) between skeletal muscle tissue and the heart tissue, and enriched genes (N = 40) shared between the liver and the kidney. This highlights the similarities due to function and/or cellular composition.

Rat to human comparison of whole-body Rat tissue RNA expression profile.

As the last part of my project, I pursued to compare the transcriptomic data of the rat and the human, which is also a key section in the mammalian atlas. To compare the two organisms, it is key to establish a new dataset. For this I established a dataset containing only ortholog genes (16,157 orthologs) based on ortholog list provided by Kale von Feilitzen from the HPA. This list is based on the human-rat ortholog list of Ensembl built 103, however filtered for high quality orthologs. The HPA list includes 14,104 one2one, 1,683 one2many and 370 many2many ortholog gene pairs.

A second step in establishing a comparable dataset is to select and pool the comparable organs. For this I created a new tissue hierarchy named comparison tissue. The expression data is pooled accordingly to Figure 6, by taking the maximum gene expression value of the lower hierarchy tissues. However, not all tissues from the rat are represented in the human dataset. Thus, from the 100 rat tissues 73 tissues were pooled into the comparison dataset. The expression dataset for cross species comparison is made up of 34 comparison tissues.

With the dataset established one can proceed to compare the gene expression profiles of both organisms. **Figure 7A** shows a dendrogram based spearman distance calculated from limma batch corrected expression values. It is visible that most comparison tissue pairs cluster together however not all of them. Some cluster rather with related tissues of their own species, which is the case for example with the female reproductive tissues: vagina, cervix, endometrium, and ovaries. Still, other comparison tissues for example tongue and parathyroid gland are rather far from each other in the dendrogram. To see another depiction of the spearman distance between all comparison tissues, one can observe the pheatmap in **Figure S4**. The testis, brain and retina tissues seem to have the furthest distance to the other tissues in terms of correlation. To further build an overview of the tissue's relationships, I visualized the expression profiles of the tissues

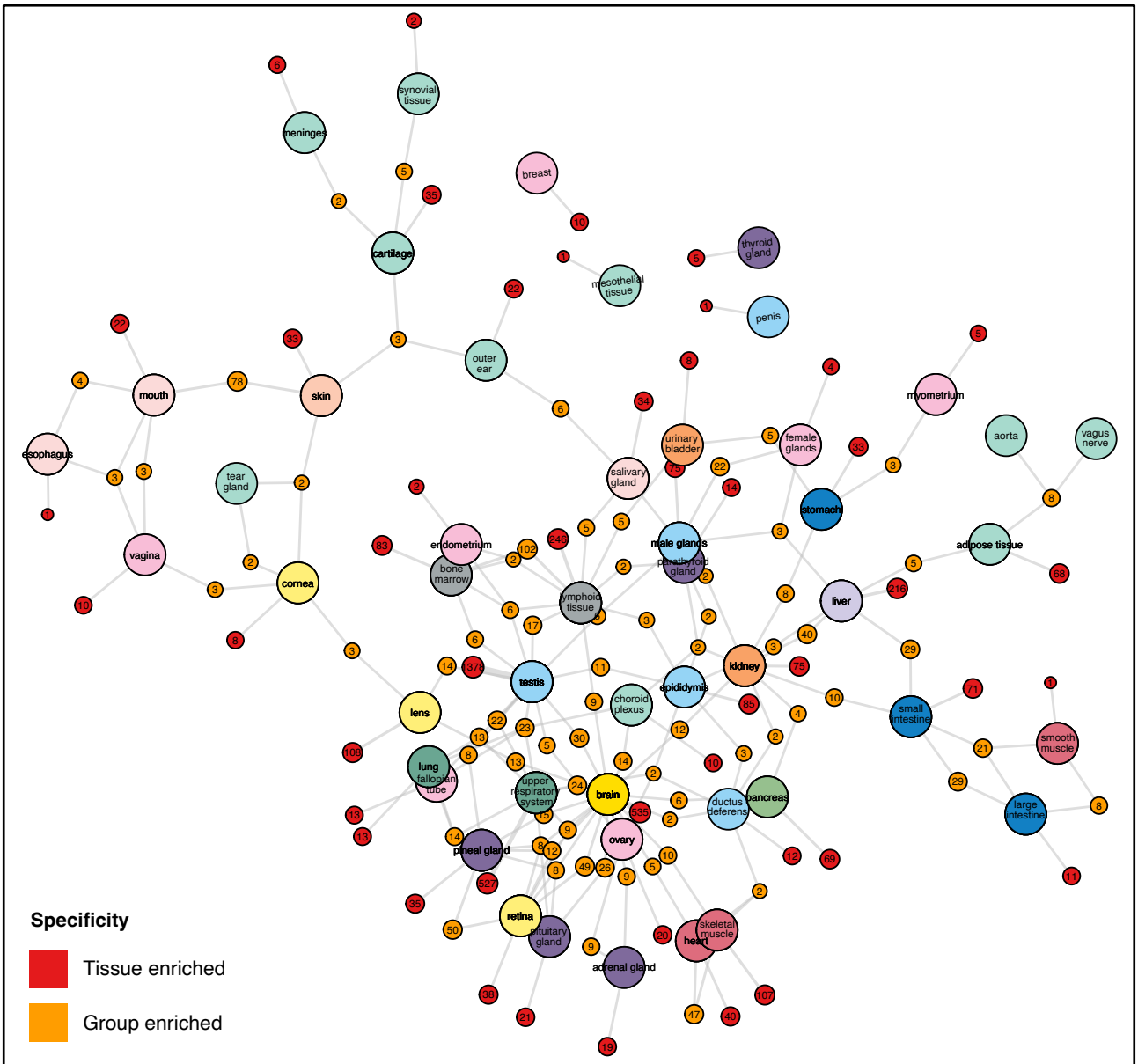

Figure 6. Network plot depicting the number tissue enriched genes for each grouped tissue, as well as the number of overlapping group enriched genes for grouped tissues. Nodes were filtered to a set of rules to reduce visual complexity. Included were only nodes, that contained tissue enriched genes, contained at least 5 genes or ranked top 2 on a selected node and contained at least genes.

through UMAP dimensionality reduction (**Figure 7B**). One can observe in the plot that most tissue of pairs cluster together, which is expected. However, it is apparent that there are some pairs that show a very different expression profile, most noticeably the tongue. The rat tongue tissue is in proximity of skin tissues, while the human tongue closer to muscle tissues. Other seemingly distant pairs are the salivary gland, urinary bladder, cervix, seminal vesicle, prostate to name a few and assessed by visually inspecting the plot alone.

The Alluvial plot (**Figure 7D**) shows that most tissue enriched genes in the rat remain enriched in human and most low specificity tissues remain having a low specificity. It is worth highlighting that around half of the tissue enriched genes classifications remained conserved between both organisms. These tissue enriched genes conserved in both amount to 1,202 or 7.4% of the ortholog genes. (This could be valuable information when investigating tissue markers or key disease actors unique and essential to a tissues function that could be studied in the rat for the development of medicaments and later applied in humans.) Similarly, 5297 or 32.8% of genes remained having a low tissue specificity. These conserved low specific genes would then possibly play a central role in cell maintenance and survival.

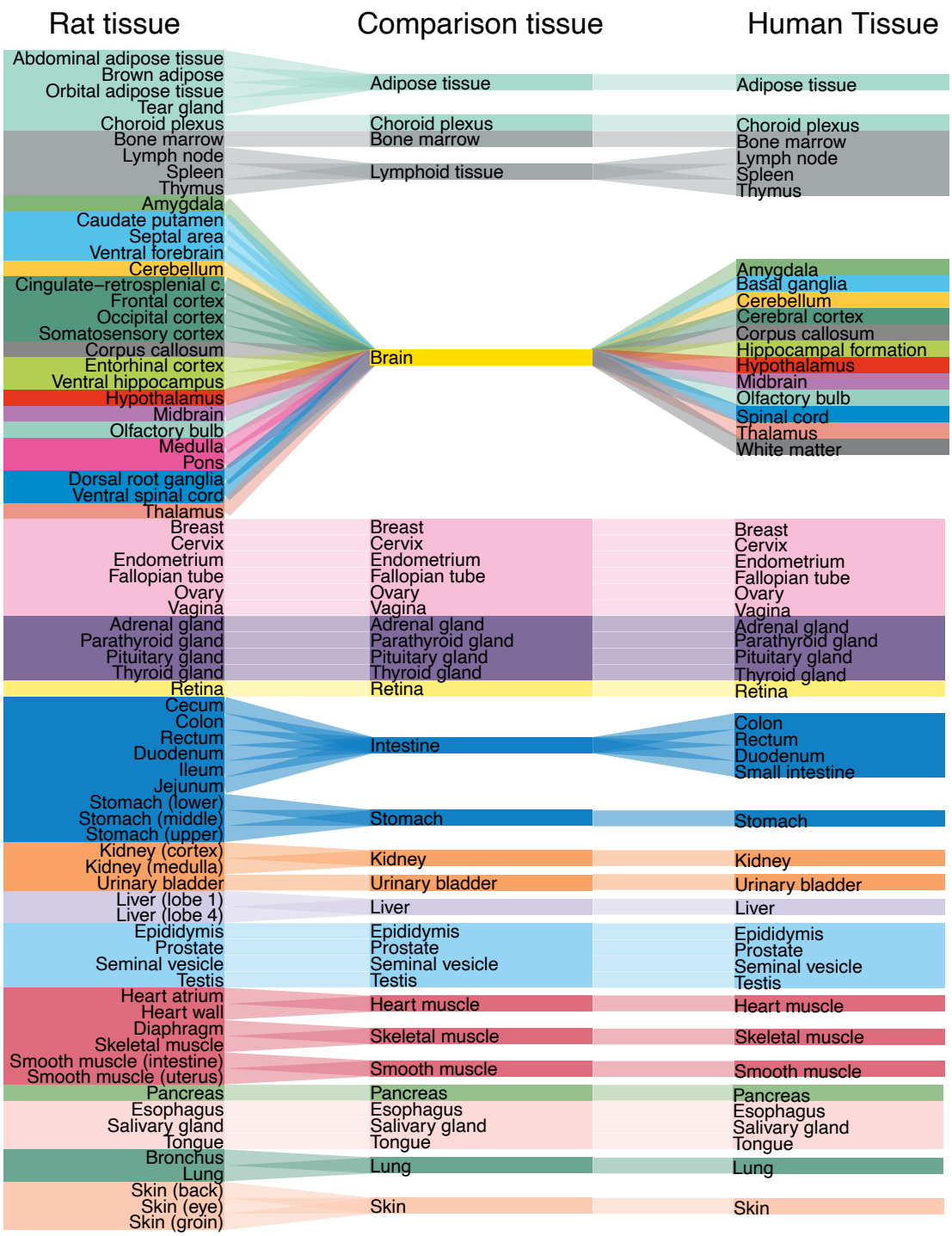

Figure 6. Comparison tissue dataset generation related to the rat tissue types on the left, and to the HPA and GTex human consensus tissues gene data. Comparison tissue dataset was generated through pooling the maximum gene expression level in nTPM of each gene of the tissues types inside each comparison tissue.

Finally, to statistically assess the similarities and differences between rat and human, a set of hypergeometric tests were performed comparing all rat comparison tissues against all human comparison tissues. (**Figure 7C**) Through this test one can assess the similarities of tissues based on the relative number of genes being categorized as elevated in both compared tissues. For example, the test shows that the rat brain tissue to have similarity in enriched genes in the human retina or human pituitary gland (or vice versa the human brain in rat retina and pituitary gland). This is rather expected since they share cellular composition. Similarly Skeletal muscle and heart muscles shows similarities in both organisms.

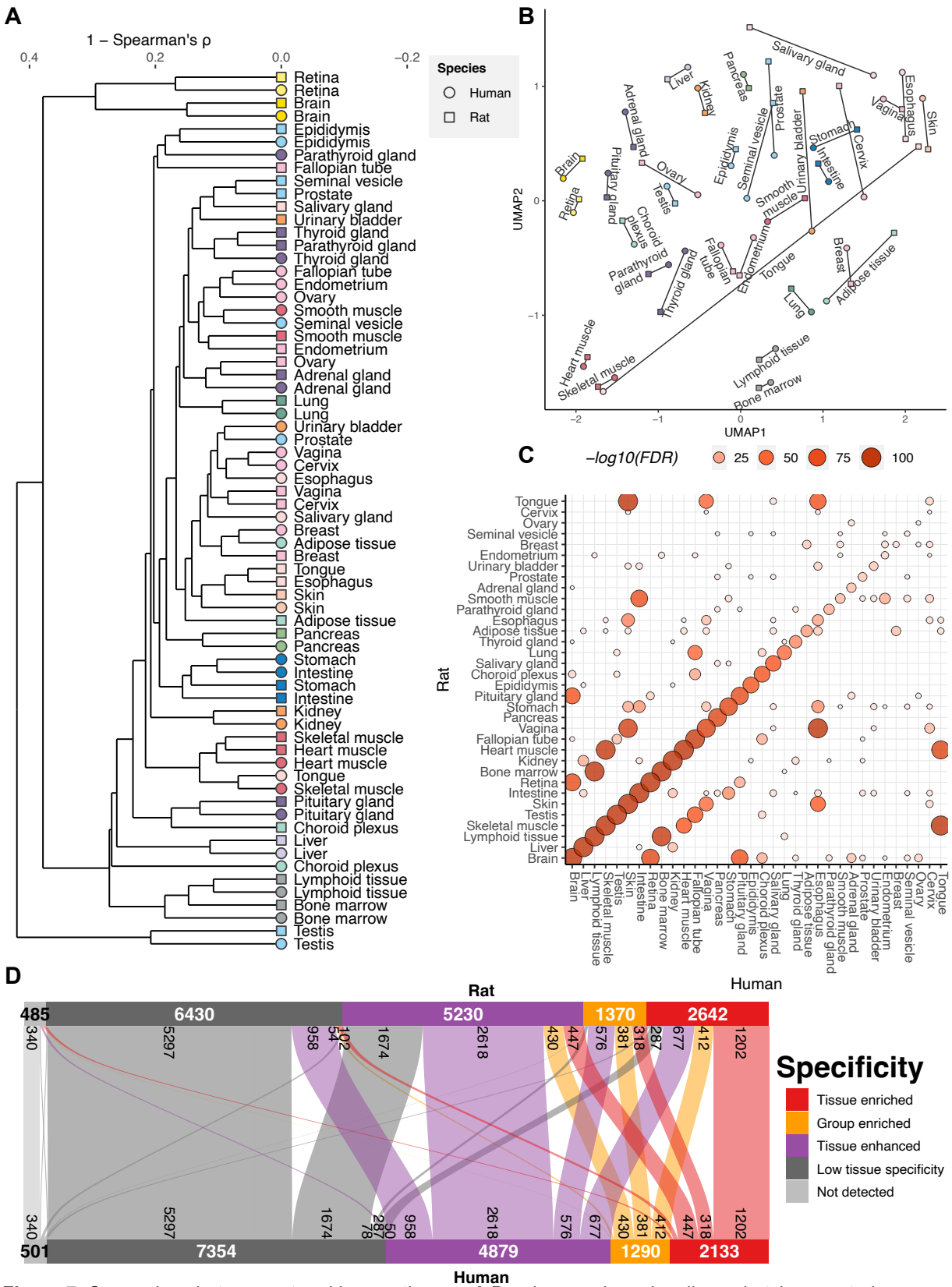

Figure 7. Comparison between rat and human tissues. **A** Dendrogram based on limma batch corrected log10(nTPM +1) values of ortholog genes of compared tissues based on sample to sample spearman distance. Dendrogram clustering calculated throughUPGMA (unweighted pair group method with arithmetic mean). **B** UMAP dimensionality reduction plot based on limma batch corrected log10(nTPM +1) values. **C** Results of the hypergeometric test assessing overlap of tissue elevated ortholog genes between all rat against all human comparison tissues. See method sections for details on the calculation. **D** Alluvial plot depicting the overlap of tissue specificity categorization between rat and human orthologs. Categorization was calculated at this comparison tissue hierarchy.

On the other corner of the plot, we find the Tongue, which does not show significance with its counterpart. However, it becomes apparent that there were sampling differences. The Rat Tongue matches heavily with the skin tissues, and the human tongue matches heavily with skeletal muscles and heart muscles. On the rat one would have possibly only sampled the top layer, while on the human tissue they might have gone deeper into the tissue.

Other interesting observations may be that the Rat vagina expression profile match with the human esophagus, which might be due to the mucosal tissue layer on both. Or a similar observation also detected on the pig atlas, where the Rat lung matches the human fallopian tube, which could be explained due to both tissues having ciliated cells.

Discussion

The Rat RNA Atlas is to be a database where one can access the gene expression profiles of multiple tissues spanning the whole body of the rat. It will complement the already available Pig RNA atlas and therefore, the data processing and analysis for this project was inspired by the work previously done for the Pig RNA atlas. They will both be integrated to form the Mammalian RNA Atlas, a sister atlas to the Human Protein Atlas (HPA).

My workflow in the project can be grouped into four steps: Normalization and dataset assembly, quality control, gene annotation and human to rat cross species comparison.

As it has been established in the HPA, the gene expression data in TPM goes through TMM normalization. **Figure S2** shows the effects of such normalization in the data on a selection of samples. It is apparent that the distribution of all samples become aligned, so that a correct comparison between tissues is possible. The effect is most notorious when comparing the pancreas samples, to other samples. TPM normalizes for transcript length and sequencing depth, but TMM normalizes for the sequencing of different tissues.

The dataset assembly was done as defined previously by the Pig RNA atlas and respecting the tissue hierarchy proposed already by in the HPA. Since both RNA atlases will be integrated, it is essential to be consistent in this regard.

There were some adjustments I had to make to the dataset, as determined through the quality control procedure. Through making a generalized combination of plots for every tissue I managed to assess the consistency of samples in the dataset and select the tissues that showed irregularities and assessed if they were sample mixups or showed contamination of adjacent tissues and made the necessary corrections.

After having implemented the changes, the resulting dataset behaves as presented in **Figure 1** and **Figure 2** and as introduced in the results section. All these figures show how the samples, and the tissues behave relative to each other in terms of variability and correlation. However, the key point, is to show that samples behave consistently across tissues and that related or similar tissues in terms of functionality, and cellular composition cluster together in dendrograms and show close spearman distance with each other relative to more distinct tissues. The dataset seems to have good and consistent quality, which allows for further analysis and the gene annotation.

Figure 3 and **Figure 4** present an overview of the annotation of genes in regards of specificity and distribution. This gene characterization is an integral part of how the RNA atlas would work as a database, enabling a quick search on tissue specific genes or for genes that are unspecific and expressed in all tissues. This would help for researchers in hypothesis generation for example, or also for gene function clustering, which could be a follow-up analysis to this project. However, a valuable feature enabled through this type of category-based gene annotation, is that one is able to quickly draw comparisons between tissues, just as presented in the network plot in **Figure 5** for example. It is hard to interpret the expression of genes through the nTPM alone, since it is far more informative to know how the gene expression in a tissue relates to other tissues. This discrete categorization allows to transmit quickly and, in a tissue-specific manner the meaning

behind the nTPM values. However, the drawback is, that the categorization is precisely discrete and based on arbitrary values. Which is why I complementing the categorizations with the Tau value, which is a scale ranging from 0 (not specific and expressed in all tissues equally) to 1 (highly specific, highly expressed in one tissue, not expressed in others). This enables for a continuous scale for specificity without being arbitrary, however the information on the tissue specific manner is lost. On the violin plots on **Figure 3C** and **Figure 4C** one can see how the discrete and continuous specificity values relate to each other. The relationship is as expected, where the most specific categories have highest tau values, while the least specific have the lowest tau values. Both methods show similar information, but they still are able to complement each other.

Besides the specificity and distribution gene annotation, another essential part of the mammalian RNA atlas would be the cross-species comparison to human expression profiles. The planned mammals to be included are as for now the Pig, Rat, Mouse, Macaque and Human. These species are known to be used in biomedical and pharmacological research, which is why it is so relevant to include this. The plots in Figure 7 present us how the expression profiles of both organisms behave with each other. As described in the results, there are most tissues are cluster together with their pair, however this is not the case for all. The exact reason behind this would partly be due to sampling differences (e.g. Tongue), however others might indeed be biological. For the exact one would need to assess the sectioning methods in for both datasets to have been done in a comparable way. Further in-depth investigation would be need to draw strong conclusions. However, by perusing this in the future, this comparison would allow us to assess better the decision on which model organism to take in drug trials and other biomedical research. It would be ideal to run experiments on tissues more similar to the human, if the end goal is to study human tissues or apply certain treatments on humans.

Additionally, one may like to focus especially on genes that are highly specific or detected only one or few tissue types, due to their uniqueness in function and probably their role in enabling the tissue to perform its biological function. However, genes that have low specificity and are detected in all, are most probably these genes that are essential for cellular function in the organism and are what some call “house keeping genes”. (**Figure 4D**). Furthermore, the comparison and overlap between the Rat and the Human gene categorization (**Figure 7D**) allows us to see that the specificity behavior of most of these genes are conserved between species. These genes would be interesting to investigate because of being potentially indispensable for an organism.

Conclusions

The rat being an established model system, information on their tissue specific transcriptomic profiles would be a valuable resource in biomedical research wherever the rat would be employed as a model animal. The Rat RNA atlas is planned to be one of other several databases collectively forming the Mammalian RNA atlas. Identification of differentially expressed genes on body-wide tissue samples and able to quantify and describe in detail the tissue specificity of all protein coding genes expressed is key to describing the rat as a model system. Based on the rat’s high use rate as an animal for scientific purposes and a lack of an accessible and whole-body spanning gene expression database for the rat, the Rat atlas would be a good complement to anybody performing experiments involving the rat in a molecular life science research environment.

Based on the quality control and data exploration done for this project, I would conclude that this dataset has good consistency and high quality to form an atlas for rat gene expression at multiple tissues. The atlas will be accompanied by gene annotations regarding tissue specificity and tissue distribution categories as well as a tau value, to facilitate a quick interpretation of the expression values while maintaining a body wide perspective on the gene’s expression. This helps us both compare a gene expression between tissues as well across species as well as to compare tissue expression profiles as a whole. The Rat RNA atlas would be a suitable resource for exploration of expression profiles of the rat and hypothesis generation for future research.

References

1. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol* **15**, (2017).
2. Strachan, T. & Read, A. P. *Human Molecular Genetics*. (CRC Press, 2019).
3. Modlinska, K. & Pisula, W. The natural history of model organisms the norway rat, from an obnoxious pest to a laboratory pet. *Elife* **9**, (2020).
4. European Comission. *Summary Report on the statistics on the use of animals for scientific purposes in the Member States of the European Union and Norway in 2019*. https://ec.europa.eu/environment/chemicals/lab_animals/pdf/SWD2019_Part_A_and_B.pdf (2022).
5. EMBLE-EBI. Expression Atlas. https://www.ebi.ac.uk/gxa/about.html.
6. Yu, Y. *et al.* A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat Commun* **5**, 3230 (2014).
7. Karlsson, M. *et al.* Genome-wide annotation of protein-coding genes in pig. *BMC Biol* **20**, (2022).
8. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (1979)* **347**, (2015).
9. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
10. 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).
11. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

Supplemental Figures

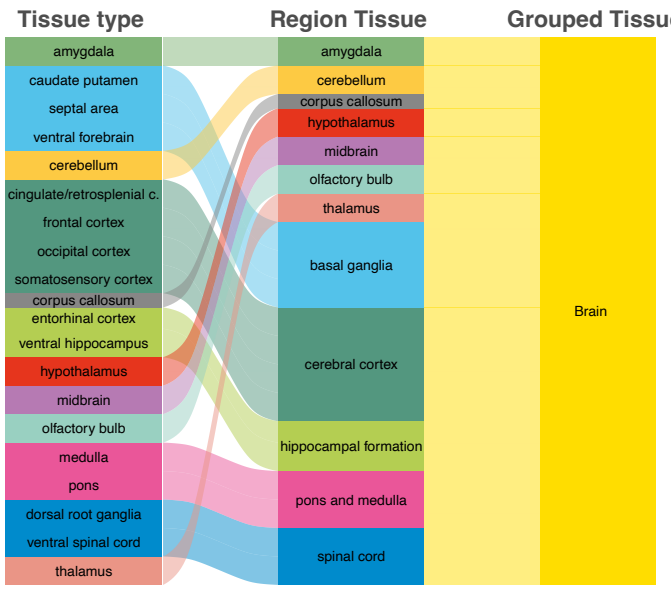

Figure S1. Representation of hierarchy order of brain tissues.

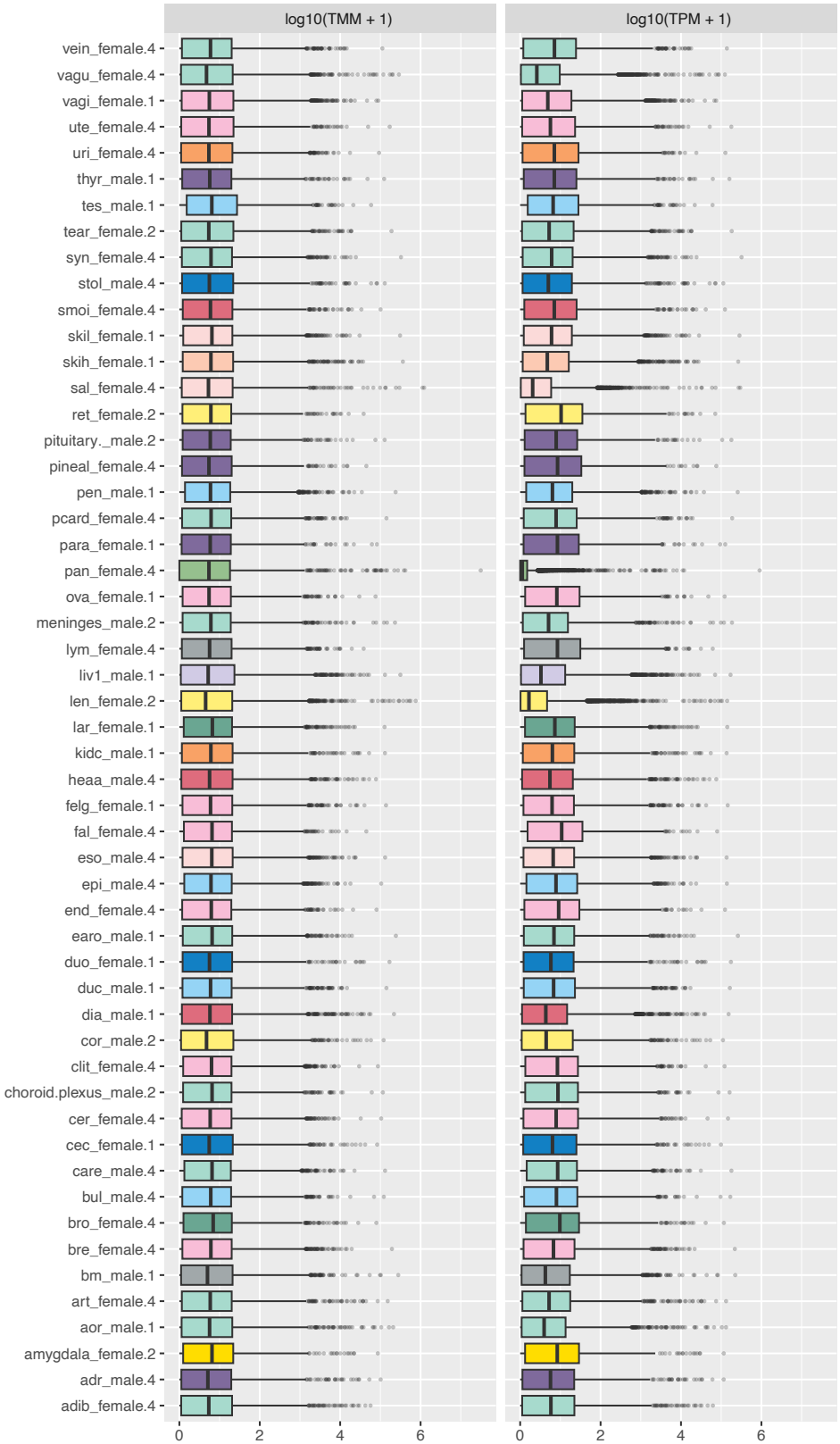

Figure S2. Comparison of different gene expression metrics between transcripts per million scaled counts (TPM) and trimmed means of M values (TMM) normalized TPM values, also referred to as normalized TPM or nTPM for short. Represented are one sample of each grouped tissue, to demonstrate the differences between both metrics. The box-plots represent the distribution of the expression levels of all genes in each sample.

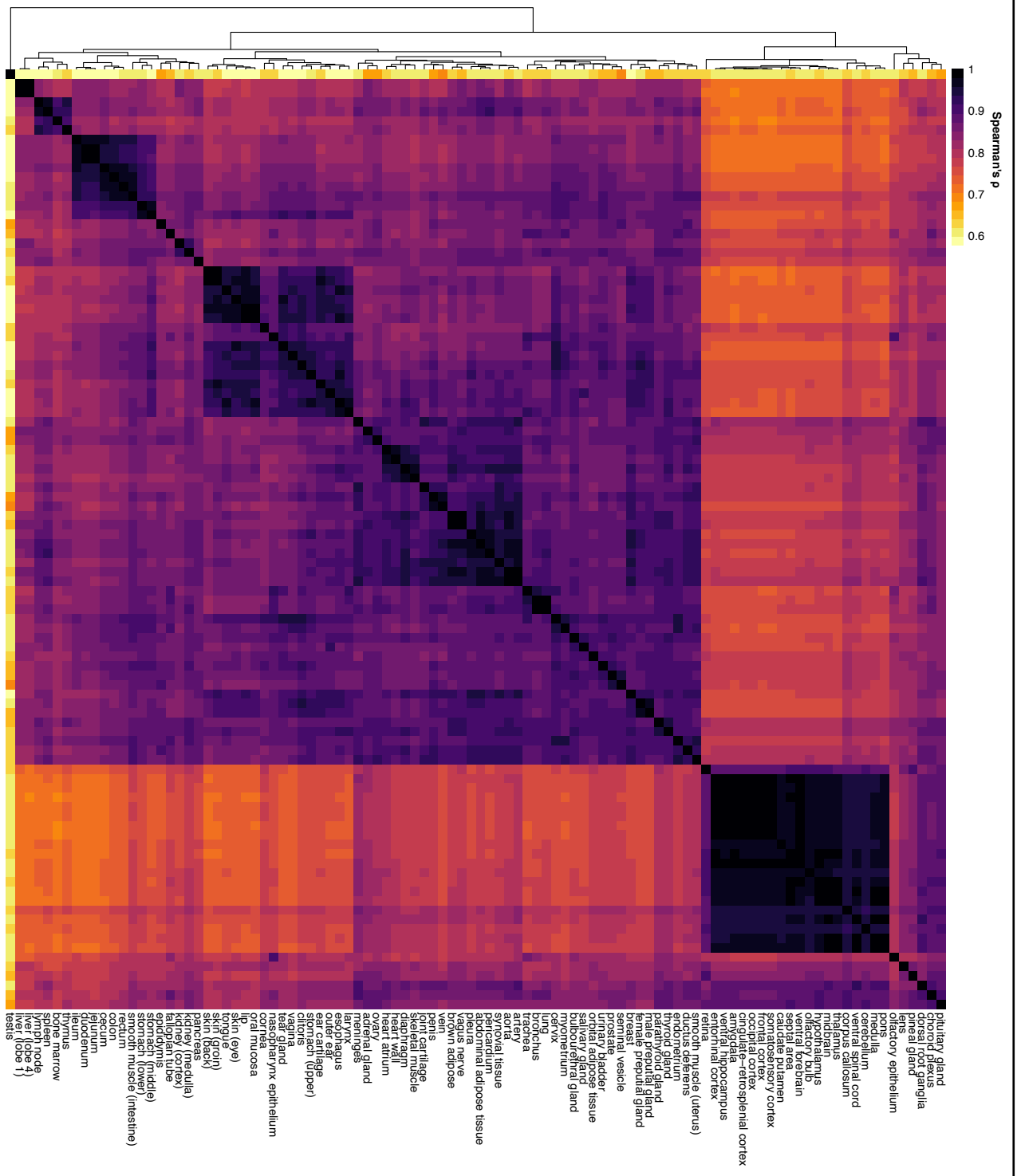

Figure S3. Clustered heatmap showing of the spearman correlation of all tissue types against all tissue types. Clustering base on euclidean distance.

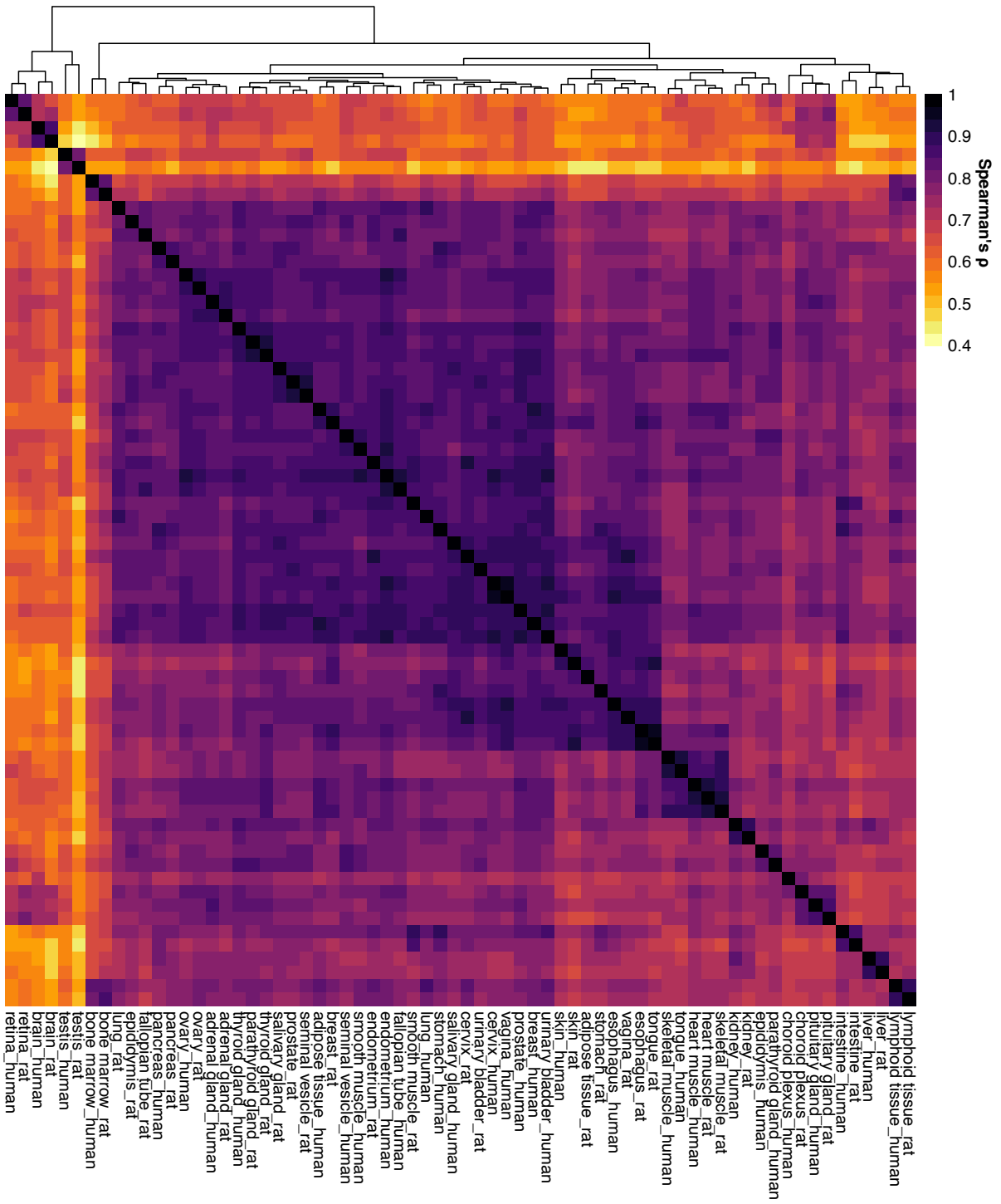

Figure S4. Clustered heatmap showing of the spearman correlation of all comparison tissue against all comparison tissues of both speices. Clustering base on euclidean distance.