Final Report

Gruppe 4

2022-07-03

Contents

1 Abstract

2 Introduction

2.1 TRAs as a tool to gain insight into embryonic development

Tissues specific antigens (TRAs) are important for the analysis of our dataset. TRAs are connected to the negative selection in thymus, which prevents an immune reaction against selfantigens. TRAs are ordered in clusters and controlled by autoimmune regulator (AIRE) in medullary epithelial cells (mTECs) and represent the diversity of antigens in our different

tissues (Dinkelacker 2019; Murphy & Weaver 2018).We use the definition that TRAs are genes, which are expressed more than 5 times the median in less than 5 different tissues (Dinkelacker, 2019). The role of TRAs is the target of this report.

2.2 Embryonic development during the observed timeframe

Our expression data was gained from embryos between week 4 to 9, hence it's essential to know what happens in organogenesis during this week in the most prominent organs: the liver, the gonades, the spleen, the heart, the stomach, the skin, the skeletal muscles and the brain. First, the liver sprout begins to form in week 3 after gestation. From week 4, it develops hepatocyte precursors and is innervated by veins. Between week 5 to 9 the production of gallic acid starts. Furthermore, glycogen granules develop by week 8 and glycogen synthesis starts in the following week (Deutsch 2013). The testis initially develop as non sex-specific gonads. At week 5, the first germ cells appear in the gonades. The gender-specific development into ovaries and testis occurs first by week 7 (Benninghoff 1993). The spleen first appears at week 6 of embryonal development. Blood vessels in the organ develop from week 8 to 9 (James & Jones 1983). Especially important is the spleen's role in the human immune system. To that end, B-lymphocytes are present within this tissue from week 12, while T-cells can be found not earlier than week 14. Previously, these cells develop in the liver starting week 9 and in the thymus from week 7 respectively (Hayward 1983). At week 3, the heart consists only of a preliminary tube. From then on the heart undergoes extensive growth and development, with cambers and ventricles forming. The fundamental layout is already present by week 5. Then, further remodeling takes place until around week 7, when the major steps are already completed (Ulfig 2009; Hikspoors et al. 2022). The stomach develops from the foregut. The primitive gut divides into foregut, midgut and hindgut by week 4. At the end of that week, the stomach is first visible (Kluth et al. 2013). Gastric pits form by week 8, while most essential cell types (enteroendocrine cells, mucous cells) appear between the 10th and 15th week and stomach acid is only secreted from week 32 onward (Esrefoglu et al. 2017). Skin development starts immediately after gastrulation at week 3. The ectoderm further develops to the nervous system and skin epithelium. There, the epidermal differentiation is illustrated through the expression of keratin genes. Adhered cells (periderm) create a protective layer for the ectoderm during weeks 4 to 8 (Hu et al. 2018). The skeletal muscles from mesoderm first in form of myoblasts that later (between week 10 and 13) fuse to form myotubes and then differentiated muscle fibers. The proteins necessary for muscle formation appear the earliest at 7 weeks, with more being expressed from week 9 and 10. Muscle fibers only form from week 15 onwards (Romero et al. 2013). While to neurulation happens around week 4, the major parts of the brain are already visible by week 9. Between theses stages, characteristic steps of neuronal development such as neuronal proliferation (starting at week 4), neuronal differentiation (starting at week 4), neuronal migration (starting at week 9), synapse formation (starting at week 9) and programmed cell death (starting at week 20) take place (National Research Council and Institute of Medicine 2009; Müller & Hassel 2018)). These processes are influenced through signals provided for instance by chemokines.

2.3 Chemokines and brain development

Chemokines are a group of small proteins, acting as chemoattractors on effector cells. They are classified in 4 groups (alpha to delta), depending on the position of their first cysteines (C). In the alpha group (or CXC), they are separated by a single aminoacid. In the beta group (or CC), they are next to each other. In the gamma group (or C), there is only one cystein present. In the delta group (or CX3C), they are separated by three aminoacids. (Yusuf et al. 2005). Chemokines induce cell migration by binding to their respective receptors (a G-Protein coupled receptor), which are often shortened with an R. For example CXCR4, the receptor of the alpha class ligand CXCL12. The function of chemokines during embryonic development is a target of further research. Nevertheless, the CXCL12/CXCR4 signalling pathway plays an important role in the neuronal cell migration (Tiveron & Cremer 2008).

3 Methods

3.1 Programming language and Libraries

The freely available programming language R version 4.2.0 and its IDE RStudio were used to draw statistical conclusions and generate informative plots. The used code packages were installed from CRAN, an online network with submitted libraries for specific programming and statistical purpose. More precisely, some packages were downloaded from bioconductor, an open software library build by developers of the community specifically for biological assays and statistical genomics. Packages for annotation purposes of microarrays were provided by brainarray.

Following libraries were used:

Table 1: *All libraries used for the code of this report, libraries installed from CRAN, Bioconductor and brainarray*

Library	Version	Library	Version	Library	Version	Library	Version
affy	1.74	AnnotationDbi	1.58	biomaRt	2.52	cluster	2.1.3
clusterProfiler	4.4.4	cowplot	1.1.1	dplyr	1.0.9	enrichplot	1.16.1
factoextra	1.0.7	ggbiplot	0.55	ggforce	0.3.3	GGally	2.1.2
ggplot2	3.3.6	ggplotify	0.1.0	ggpubr	0.4.0	ggrepel	0.9.1
ggsci	2.9	ggupset	0.3.0	grid	4.2.0	gridExtra	2.3
gt	0.6.0	gtExtras	0.4.1	hexbin	1.28.2	hgu133plus2hsenstcdf	25.0
hgu133plus2hsenstprobe	25.0	igraph	1.3.2	kableExtra	1.3.4	limma	3.52
magick	2.7.3	magrittr	2.0.3	org.Hs.eg.db	3.15	pheatmap	1.0.12
png	0.7	Rcpp	1.0.9	RCurl	1.98	readxl	1.4
rentrez	1.2.3	Rfssa	2.0.1	stringr	1.4	svglite	2.1
tidyverse	1.3.1	treemapify	2.5.5	VennDiagram	1.7.3	viridis	0.6.2
vsn	3.64	webshot	0.5.3	XML	3.99		

3.2 Dataset

We obtained the data set from Yi H *et al.* (2010). We chose this data set by the following criteria, it contains human embryonic data and it covers every week between the 4th and 9th week, which are interesting stages of embryogenesis and organ development. Three replica at each point in time were tested, hence data from 18 embryos were acquired. The timezone covers the Carnegie stages 10-23, finishing the process of embryogenesis and organogenesis. This peroid of embryogenesis is highly regulated with considerable differential gene expression. Overall, the data set suits the requirements for our purpose.

3.2.1 Affymetrix U133 plus 2.0 human GeneChip array

The data was generated from embryos by using Affymetrix U133 plus 2.0 human GeneChip arrays. RNA microarrays are slides coted with oligonucleotides as matrices which screen for thousands of transcripts. The HG-U133 Plus 2.0 allows the detection of about 50,000 transcripts and uses quality control matrices. The Affymetrix chip include 62 control transcripts, whose intensities are imported together with the acquired data.

3.2.2 Importing the data set

We downloaded the raw data to a local harddrive from the Gene Expression Omnibus with the Accession Number of GSE15744. We imported it with the help of the library *affy* and is connected to the correct Annotation by the brainarray package. The *affy* package allows more manageable data analysis and manipulation of microarray intensity values.

To access the data remotely, we uploaded it to the cloud-based repository hosting service github. It can be imported with the library *Rfssa*.

3.2.3 Quality control of the suface images

To ensure that the microarrays are without surface damaged, we checked their images. We selected two images as an example

As shown in Figure 1A as an example, the surface of the chips are visible and show no spatial artefacts, fingerprints, irregular dye or stripes. Some differences in overall brightness are visible but marginal.

3.2.4 Quality control of RNA Degradation

We can further analyse the quality of the microarrays by checking for low RNA quality chips. Coted matrices degrades under unfavorable conditions, which negatively affects raw intensities (Fasold & Binder 2013). By plotting the RNA degradation for 3'–5' strand, we can compare the different chips (Figure 1B).

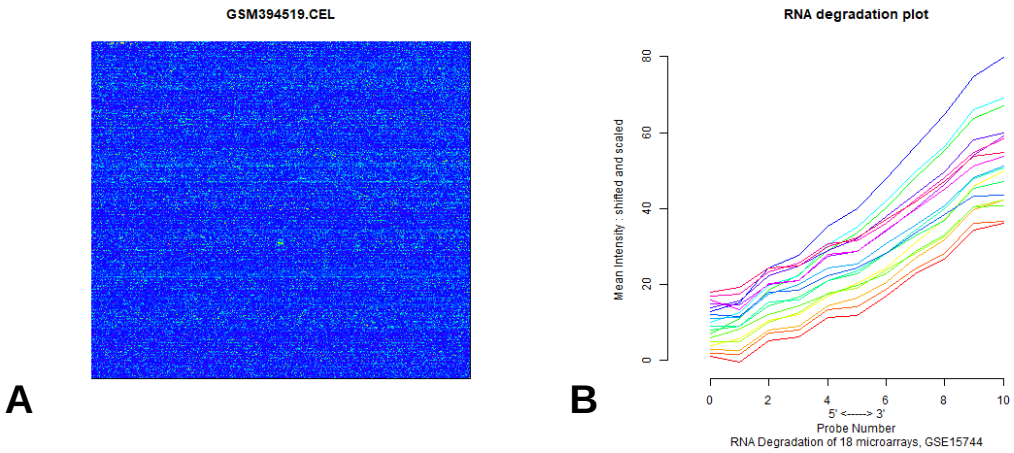


Figure 1: Quality control: Selected surface image of a microarray shows no damage or artefacts and RNA degradation plot shows slight irregularities and verifies the data. A*: The microarray inspection shows no irregularities and every chip is accepted for further data analysis.* ***B****: Some crossing lines can be seen, especally the microarray GSM394519. We decided that the inconsistencies are minor though, and kept all microarrays to avoid the loss of potentionally relevant data.*

3.3 Normalising the data set

Intensity values of different chips are affected by sstatistical variance and random fluctuation. To access the biological relevant variation the raw data is normalised. We chose the vsn rma normalization with its library *vsn* according to Huber *et al.* (2002). This library is designed to process microarray intensity values. It calibrates data and applies *generalized log*-transformation, which is an adjusted natural logharithm and preserves statistical significance.

3.3.1 Quality control of the vsn normalization

To verify the transformed data intensity values, some test can be performed (Figure **??**). After the normalization, the rank of the mean of the intensity values and their standard deviation should not correlate. Therefore we can plot the rank of the mean against the standard deviation to control the normalization method and should get a horizontal line indicated in red (Figure **??**A). Another way to control the normalization is to visualize the intensity values. Here we have two options. We used boxplots to compare each of the 18 microarray separately by its mean, median and variance. This allows us to knock out unfitting arrays (Figure **??**B). The second option gives us the ability to zoom in even further. The intensity levels of three replica should be the same, since they were taken at the same time. We can use scatterplots to compare single intensity levels. With one of the replica applied on the x and y axis respectively, we should see a scatterplot following the linear function y = x since the same transcript should show the same intensity in both replica (Figure **??**C).

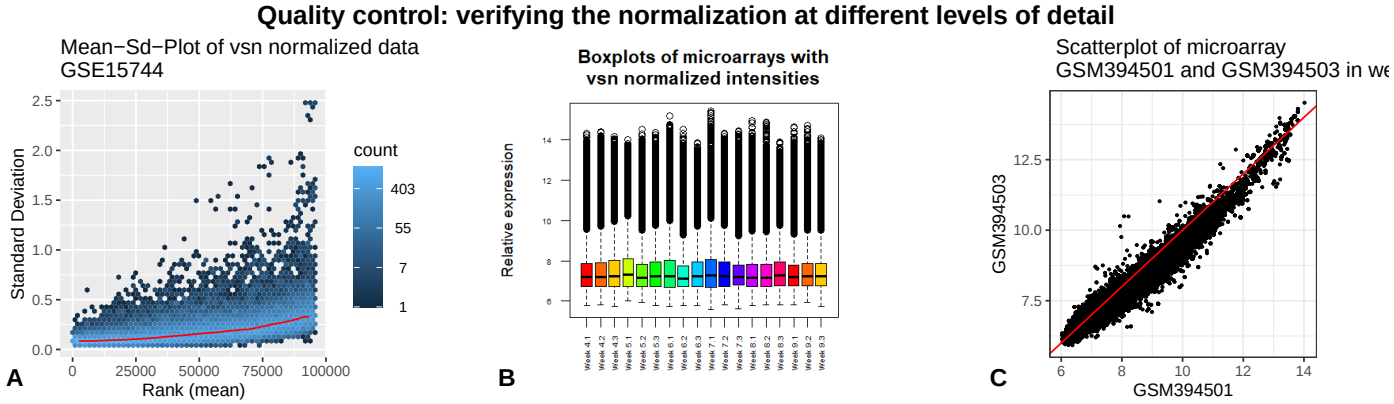


Figure 2: *The plots support the use of the vsn normalization on our data set. A: The red line is close to horizontal, although it shows some correlation at high intensity leves. B: The boxplots show nice allignment of the mean intensity values. Some outliers are given but can neglected given the 0.25 and 0.75 quantile. C: A selected Scatterplot is shown. Very slight banana shaped structure can be seen, but only marginal. Overall the quality control confirmes successful vsn normalization*

3.4 Annotation

To make sense out of the intensity values they need to be associated to common data with known properties. We applied the data frame *ensembl_103.txt* provided by Dr. Dinkelacker, to annotate our data and yield the appropriate transcript ID for the Probe ID of the microarray. To annotate for TRAs, we applied another data frame by Dr Dinkelacker called *tra.2017.human.gtex.5x.table.tsv*.

3.5 Limma package

The *limma* package determines among many other things the changes of gene expression over time in intensity values of microarrays. It facilitates advanced statistical algorithms to calculate the necessary coefficients of a linear model for every intensity value in the data set. It uses information borrowing, quantitative weighting, variance modelling and data preprocessing, while not subset the data (Ritchie. et al. 2015). Because the linear model was casted on every intensity value, statistical tests called Empirical Bayes can determine differential expressed genes via t-statistics and their associated p-values.

3.6 Over representation analysis

The statistical method over-representation analysis determines among other thing the over represented function of genes with associated transcripts in a subset of a mother data set with annotated transcripts and known functions. Categories for functions can be accessed via gene ontology.

4 Results

4.1 Limma analysis

To filter our data for biological interesting data, we performed *limma* analysis to extract differentially expressed genes. Our threshold for significance is an Benjamini-Hochberg adjusted p-value of 0.01 or below. We found changes of gene expression in 1,814 transcripts.

4.1.1 Volcano plot

The gathered dataset from limma analysis of the differential expression between weeks 4 to 9 was used to created a volcano plot. The negativ log10 of the adjusted P-value was ploted against the logFC value. The -log10 (adjusted P-value) boundary was set at 2 which equals our targeted adjusted P-value of 0.01. The logFC boundarys were -1 and 1, which reflects a doubling in the expression (Supplementary: Figure 9). The genes were categorized in two groups:

1. up-regulated if the logFC value is larger than 1 and the adjusted P-value is smaller than 0.01
2. down-regulated if the logFC value is smaller than 1 and the adjusted P-value is smaller than 0.01

Genes which didn't belong in those groups were declared as not differentially expressed. The differentially expressed genes between all weeks with an adjusted P-value smaller then 0.01 were further used.

4.2 Comparison of differential expressed genes between the original paper and our method

We acquired the data from the original paper from the supplemental materials. We imported it into a data frame annotated it with the use of our data set because the both have the same probe set IDs from the same microarray.

4.2.1 Venn diagram

To determine the number of transcripts which overlap between the data of the paper and our data, we plotted a Venn diagram (Figure 3).

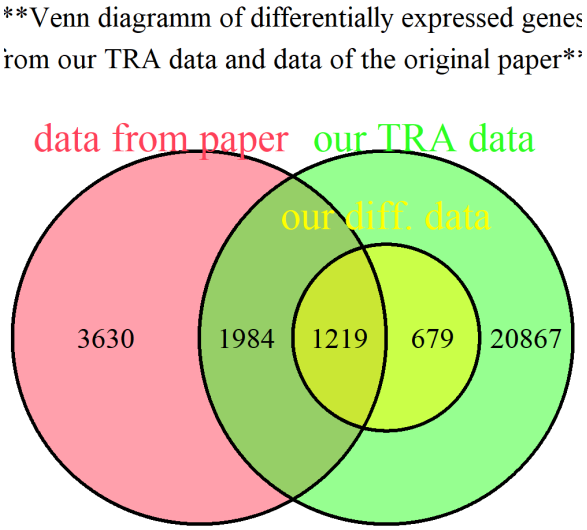


Figure 3: The Venn diagram shows overlapping sections between our data and the data from the original paper. *According to the paper, of the its 6,833 transcripts are differentially expressed. This overlapps with our general data with all TRAs (24,749 transcripts) and with our differentially expressed (diff.) data according to our limma analysis with a p-value of 0.01 (1,837 transcripts)*

The intersection showed a number of transcripts worth for further analysis.

4.2.2 Verifing the trends postulated in the paper with our data

In contrast to our method using limma, the authors of the original paper used *One-way analysis of variance* with a p-value of 0.05 to determine the differential expressed genes doi: 10.1096/fj.10-158782. In the paper they provided an annotation of transcripts that were regulated *up*, *down* or showed an *arch*. We used k-means clustering to determine, if we see these trends in our data aswell (Figure 4).

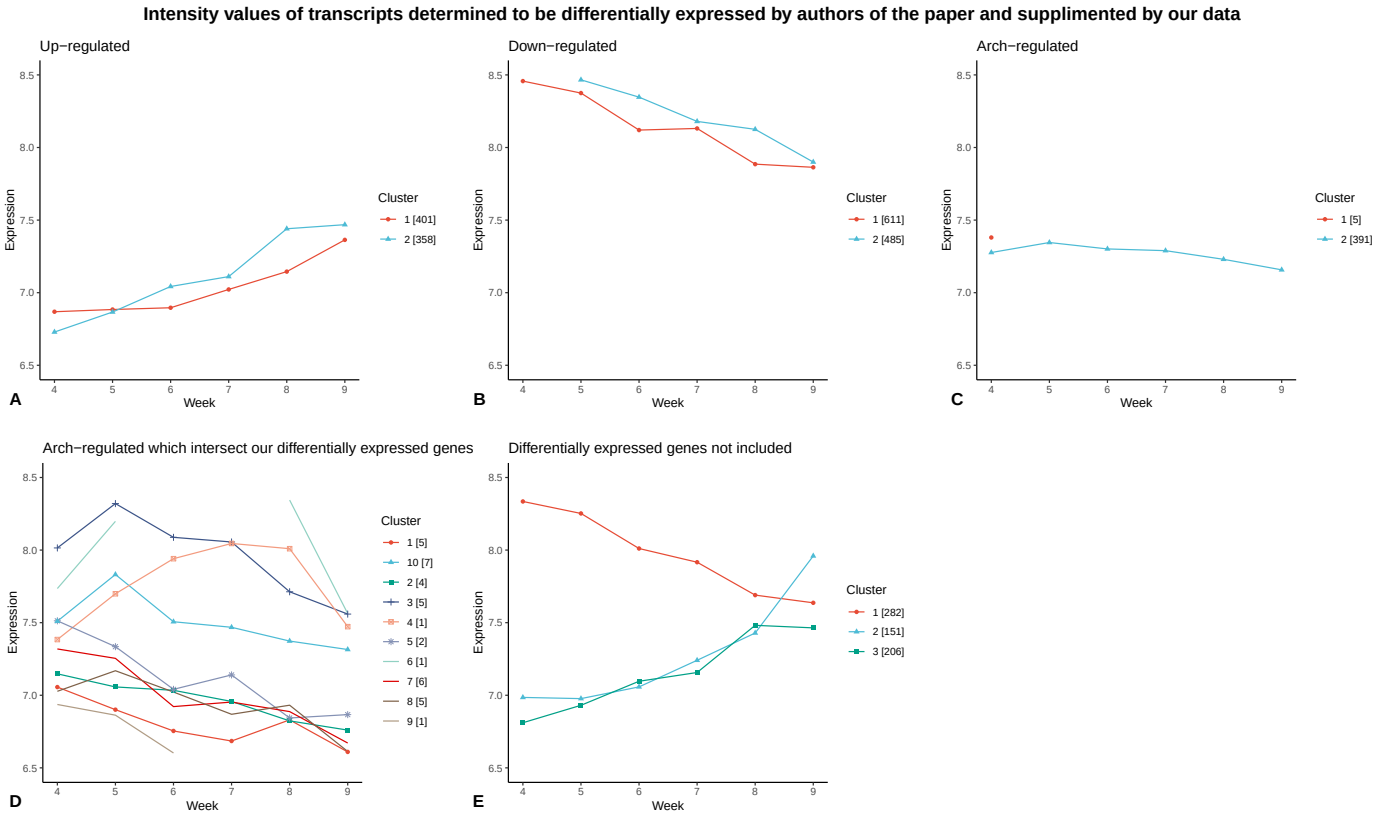


Figure 4: Clustering of the common transcripts reveals similar trends and but our additional data shows missing differentially expressed genes. *We chose data with transcript IDs both in the data set of the original paper and in our TRA data. Additionally they had to be annotated as* up, *down*, or *arch* by the authors of the paper. The* up *and* down *regulated transcript-data (**A** and **B**) follow the same pattern as postulated in the paper. For the* arch *regulated transcript-data (**C**) we see clusters, with same matching the* arch *pattern and some who are rather undetermined. Of 396 transcripts only 39 of our differentially expressed genes by* limma *analyis share this* arch *property (**D**). Here we clearly see differential gene expression. In **E** we plotted data of our differntially expressed genes. There are* up *and* down *regulated genes determined by our* limma *analysis, that are not included in the data set of the authors of the original paper*

We repeated the Figure4A and B with our differentially expressed genes aswell and the results were highly similar. The Figure can be found on our github (Report/Comparison with paper differentially expressed.png).

For Figure 4C cluster 1, 6, 7, 8 clearly show an *arch* regulated pattern, but for clusters 4 and 5 the data rather appears to be *down* regulated, or *up* regulated in the case of cluster 10. And the final clusters 2, 3 and 9 with a total of 156 transcripts are not differentially expressed at all. When we cluster our differentially expressed genes annotated with the *arch* pattern (Figure 4D), we recieve only 10% of the amount of transcripts which are clearly clusters.

Figure 4E reveals, that there clearly *up* and *down* regulated genes of 646 transcripts or about 10%, that are missed by the method used by the authors of the paper.

4.3 TRAs can infer a basic timeline of organ development

4.3.1 Differentially expressed transcripts can be linked to all analyzed tissues

The TRA Data covers 53 distinct tissues. For all of those, we found at least 40 differentially expressed transcripts within our dataset. The minimum was found with 46 stomach-linked transcripts, the maximum were 837 TRAs for the testes.

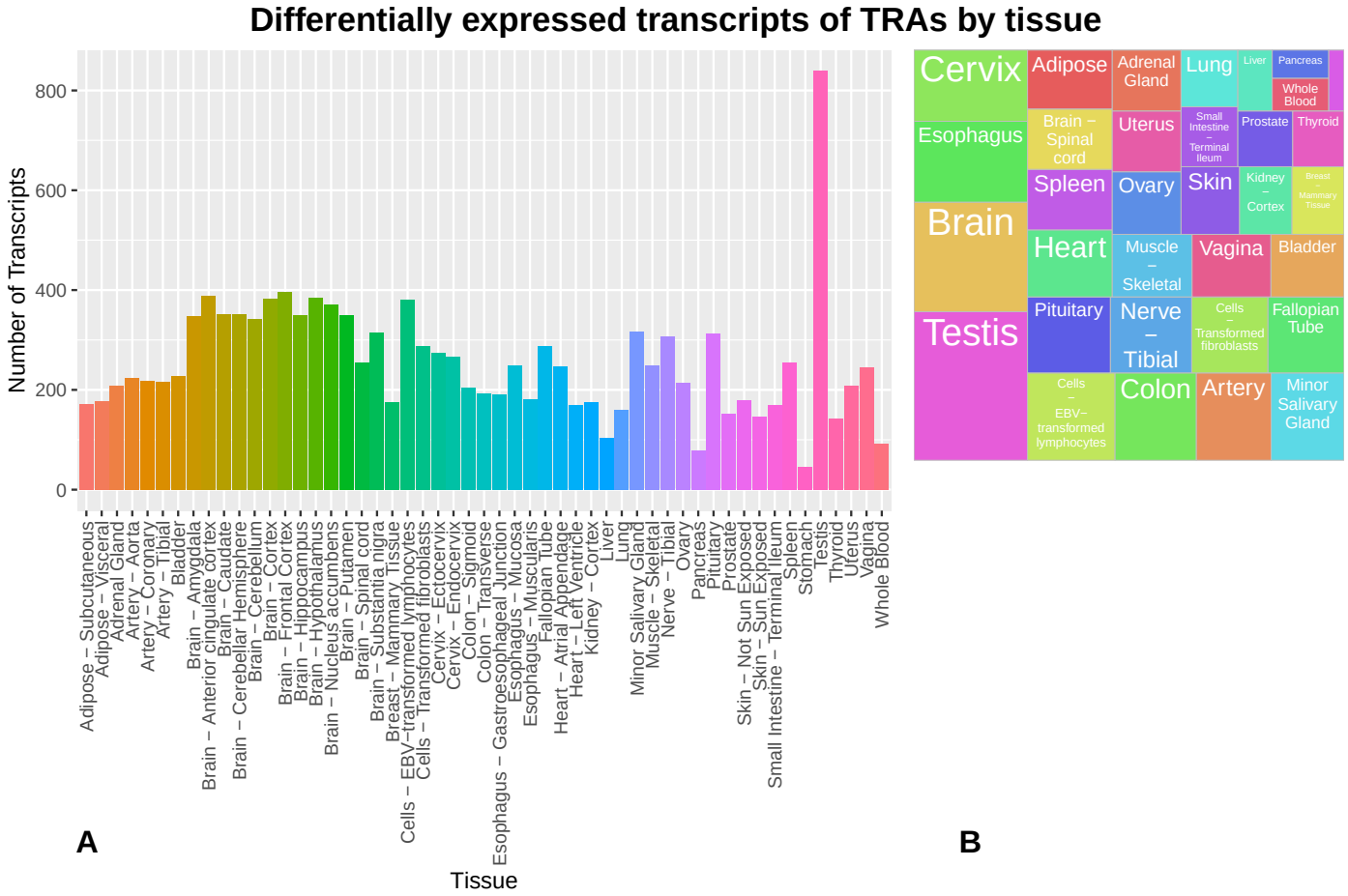


Figure 5: *A. The number of transcripts associated with each tissue, including subtissues, is displayed. B. The share of TRAs associated with each tissue. Subtissues are subsumed under their main tissue.*

```
[1] 7.403067
```

These numbers are sufficient for further analysis of the gene expression within individual tissues. Therefore, all further analysis will be based on our dataset with differentially expressed genes from limma analysis. Nonetheless, it should be noted that there is a significant overlap between the TRAs associated with different tissues, especially as each transcript is on average linked to 7.4 different sub-tissues or tissues. This overlap is further illustrated by Fig. 6.

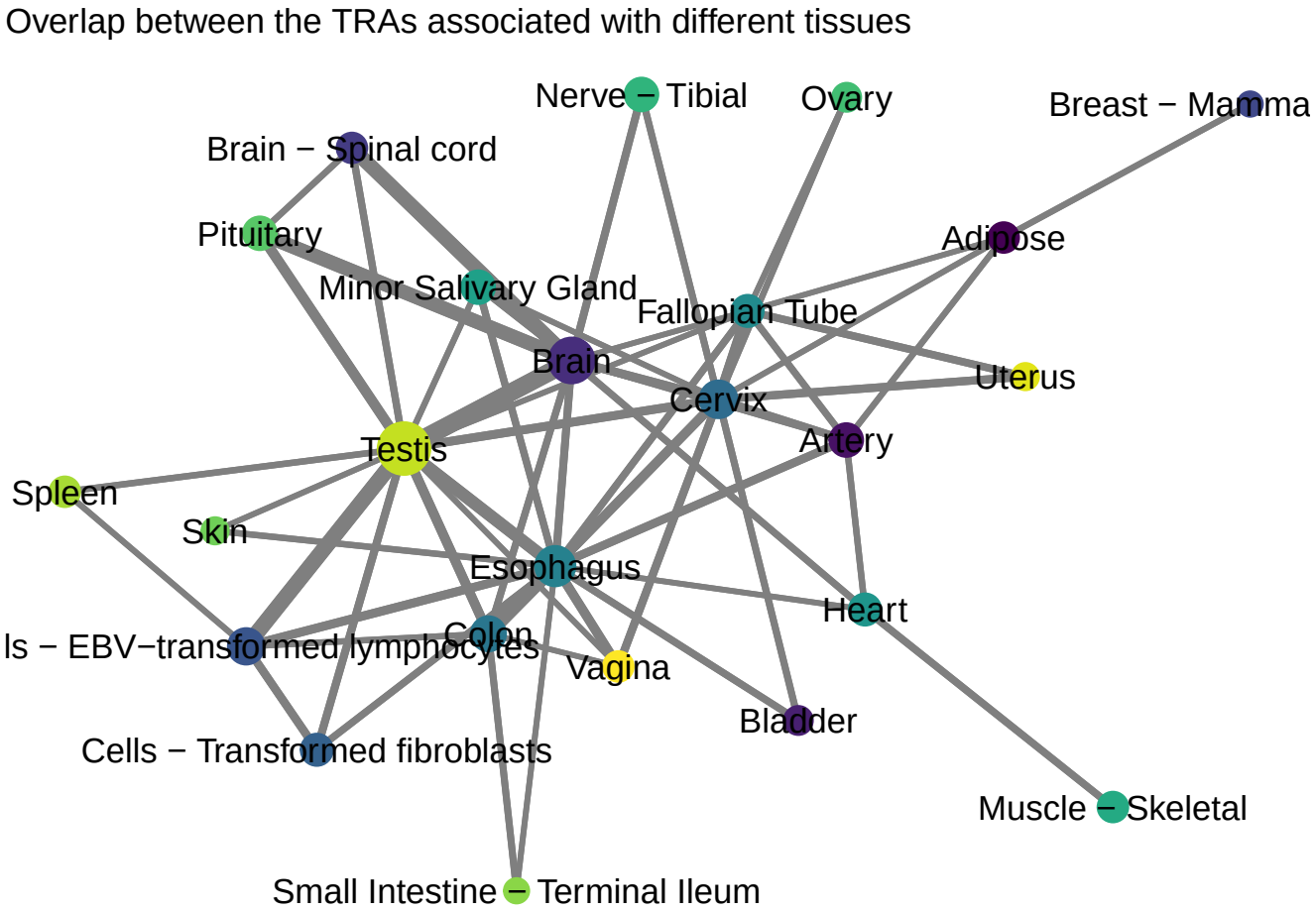
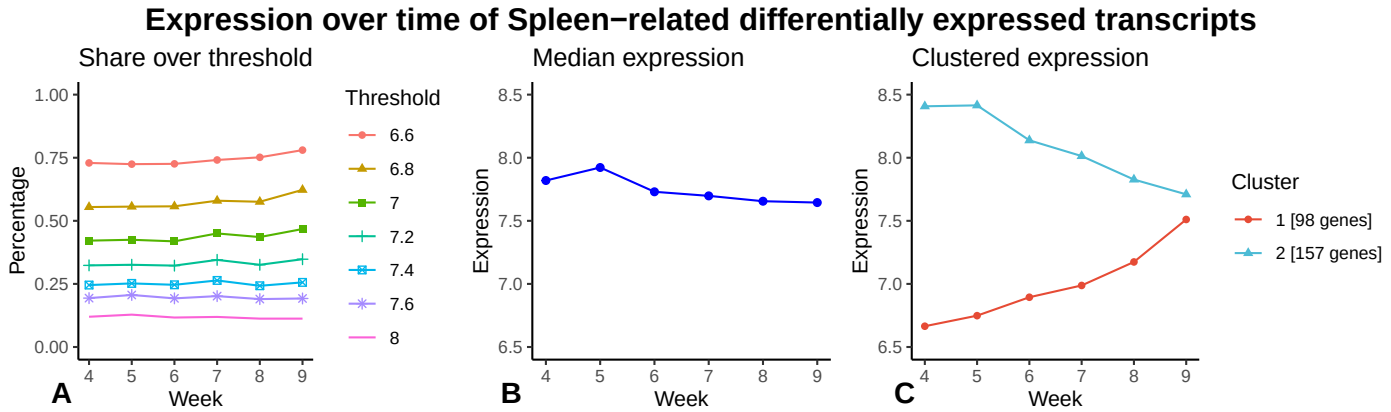


Figure 6: *Each tissue is displayed as a node with its size representing the number of transcripts associated with it. The edges show the shared TRAs between the linked tissues, with only links corresponding to more than 100 common TRAs visible.*

4.3.2 The expression of all TRAs associated with a tissue cannot be used to infer organ development

In this research, we attempt to draw conclusions about the developmental state of a tissue based on the expression of genes associated with it alone. Therefore, we analyzed the share of differentially expressed transcripts above a certain expression level over time, as shown in Fig. ???A. Furthermore, we observed trends within the median expression of all differentially expressed transcripts associated with a tissue (Fig. ???B). Since both metrics only showed in miniscule changes, we hypothesized that distinct, counteracting trends in expression exsisted within one tissue. Thus, k-means clustering was used to determine groups of TRAs with similar expression patterns. For each of these clusters, the median expression was plotted as shown in Fig. ???C.



For many tissues, as shown here exemplary with the spleen, the clustering revealed two or more clusters that could each be characterized as either an upregulation or a downregulation. In order to analyze the indications for organ development, we

analyzed the functions of the transcripts belonging to the two clusters.

4.3.3 The clusters of up- and downregulated transcripts can be linked to distinct gene functions.

For all differentially expressed spleen-associated transcripts, we used the NCBI gene database to get a functional annotation.

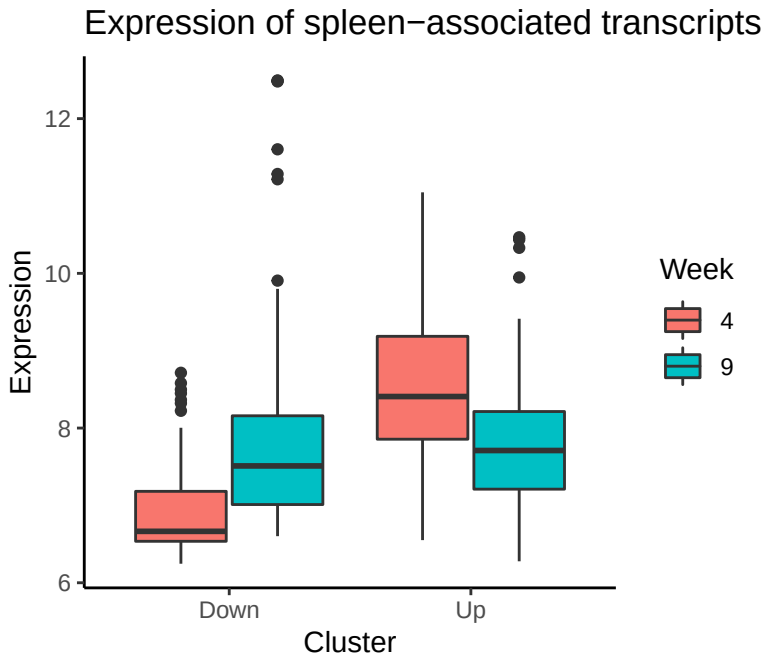


Figure 7: *A further look on the expression of transcripts in the up- and downregulated clusters shows that the upregulated transcripts are close to the minimum expression level between 6 and 7 in week 4 and showing expressions between 7 and 8.5 by week 9. In contrast, the downregulated genes have very high expression levels (8-9) by week 4 and decrease to a more moderate expression between 7 and 8.5 analogous to the upregulated transcripts.*

As shown in Fig. 7, the spleen is a clear example of two dinstict clusters with one consisting of upregulated previously inactive genes and one with downregulated highly active genes. For all these differentially expressed transcripts, we used the NCBI gene database to get a functional annotation. Of the 98 upregulated genes, 48 had a functional annotation. 17 of those were clearly associated with immune system or blood functions and thus relevant for the functional thymus. We further found 157 downregulated genes. There, 70 were annotated and 45 of those displayed a relation to the cell cycle or cell division. The tables of the transcripts with a relevant function are visible in [Suppl.].

4.3.4 Overrepresentation Analysis can create plots that signify organ development

For this analysis, the eight tissues with the most meaningful results were chosen.In Fig. @ref(fig: ORA-plot), the most important functions for these tissue were determined through overrepresentation analysis. In addition, the Expression of the associated transcripts was plotted.

For the spleen (Fig. **??**A), we determined a largely constant expression of immune-related genes throughout the time frame, with a slight increase in some functions from week 7-9. The brain (Fig. **??**B) showed an increase in neuron projection morphogenesis from a previously inactive state (expression < 6.8) in week 4 to a significant expression (>7.5) by week 8. Synaptic signaling stayed at relatively constant expression levels. Heart-associated functions (Fig. **??**C) can be grouped into two categories. Cardiac functions (cardiac muscle tissue development, heart contraction) are highly expressed in week 4 and fall continously until week 8. In contrast, general muscle gene sets are rising from originally lower expression levels during the observed time. The liver (Fig. **??**D) shows no clear expression patterns, with some metabolic functions increasing through time (organic hydroxy compound metabolic process) while others stay mostly constant (cellular amino acid metabolic process) or fall (organic acid catabolic process). In contrast, skeletal muscle gene sets (Fig. **??**E) show a very clear trend. After a mostly slight increase between week 4 and 8, a sharp rise in expression levels is visible from week 8 to 9. The testis-associated sets (Fig. **??**F) continuously decrease in expression from week 5 onward. For the stomach

(Fig. **??**G), we found a initially high expression in week 4 that then falls until week 6 and then increases again towards week 9. Finally, the skin-associated functions (Fig. **??**H) all displayed a constant rise in expression levels from week 5 to 9.

4.4 Hypothesis: Neural TRA expression patterns reflect morphological brain development

The annotated dataset after limma analysis was filtered to determine which genes are differentially expressed in a certain tissue. The discovered genes were examined with NCBI to determine their function. The genes of interest are categorized in three groups:

1.Genes of Ion channels 2.Genes for neuronal development 3.Genes of cytokines

4.4.1 Ion channel

```
integer(0)
```

```
integer(0)
```

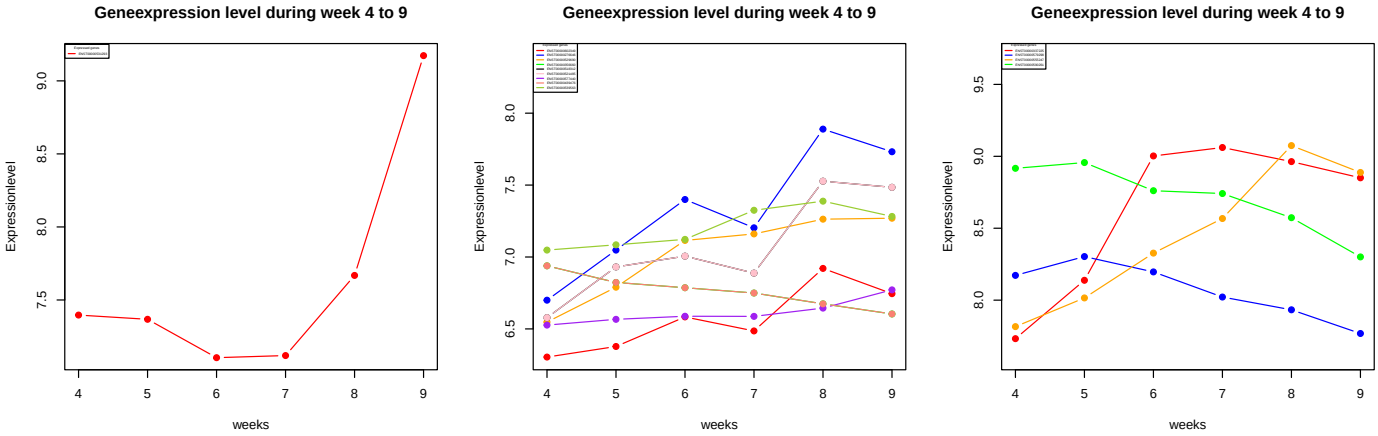


Figure 8: *Geneexpression of different genes for ion channel (left), neuronal development and function (middle) and cytokine (right) was plotted for week 4 to 9.*

4.4.2 Ion channel

Ion channels play an important role in the function of neurons. We discovered that ENST00000531293 is highly expressed in Nucleus accumbens. It shows a significant increase between weeks 7 to 9 and codes for SLN sarcolipin which is a Sarcoplasmic reticulum Ca(2+)-ATPase.

4.4.3 Genes for neuronal development and function

The second group are genes with a specific function in the neuronal development and function.

4.4.3.1 Genes for neuronal function Therefore we discovered that ENST00000276646 and ENST00000529690 show an significant increase in gene expression over the weeks. Both genes were connected to the Cerebellar Hemisphere and were associated with SYBU (syntabulin). SYBU plays an important role as it contributes to activity-dependent presynaptic assembly in neuronal development.

4.4.3.2 Genes for axon guidance Filtering mentioned four genes for axon guidance:

ENST00000602349 codes for NXPH1 (neurexophilin 1) which forms a tight complex with neurexins. Neurexins promote the adhesion between axons and dendrites. ENST00000602349 shows a strong increase, especially between weeks 7 to 8 and is connected to Anterior cingulate cortex.

ENST00000518312 and ENST00000521485 encode for SNAP91 (synaptosome associated protein 91) which plays a role in regulation of clathrin-dependent endocytosis. Therefore SNAP91 is important for axonal functions of neurons like postsynaptic density, which is essential for functional neurons (Overhoff et al. 2020). ENST00000518312 & ENST00000521485 were associated with Cerebellar Hemisphere and also show significant increase between weeks 7 to 8.

In addition ENST00000539563, encoding for LSAMP (limbic system associated membrane protein), plays a role in axon guidance. The encoded preprotein is processed into neuronal surface glycoprotein which interacts as an adhesion molecule during axon guidance and neuronal growth in the developing limbic system. ENST00000539563 is associated with Putamen which is grouped in basal ganglia. Basal ganglia were associated with the lymbic system.

4.4.3.3 Genes for neuronal survival ENST00000356660 and ENST00000439476 code for BDNF (brain derived neurotrophic factor). A binding of BDNF to its receptor promotes neuronal survival. Both genes show an identical decline in expression over the weeks. Nevertheless, ENST00000356660 is connected to Cerebellar Hemisphere and ENST00000439476 to Hippocampus.

4.4.3.4 Brain associated gene Finally ENST00000577440 encoding for SEPTIN4 (septin 4) was identified. SEPTIN4 may regulate cytoskeletal organization. A defect in septin function disturbs cytokinesis. ENST00000577440 is connected to Cerebellar Hemisphere and shows an increase between weeks 7 to 9.

4.4.4 Genes for cytokine

Cytokines are important signaling molecules. They take up an important part in the signaling process during neuronal development (Tiveron 2008).

4.4.4.1 Interleukin related genes Two interleukin related genes were discovered. ENST00000555247 encodes for IL11RA (interleukin 11 receptor subunit alpha), which is a receptor for the cytokine Interleukin 11. The IL-11 receptor is a member of the hematopoietic cytokine receptor family. ENST00000555247 is connected to the spinal cord and shows a significant increase in expression during weeks 4 to 8. ENST00000590261 encodes for LF3 (interleukin enhancer binding factor 3) and forms a heterodimer with a 45 kDa transcription factor. This complex is necessary for T-cell expression of interleukin 2. ENST00000590261 is related to Substantia nigra and shows a decline over the weeks.

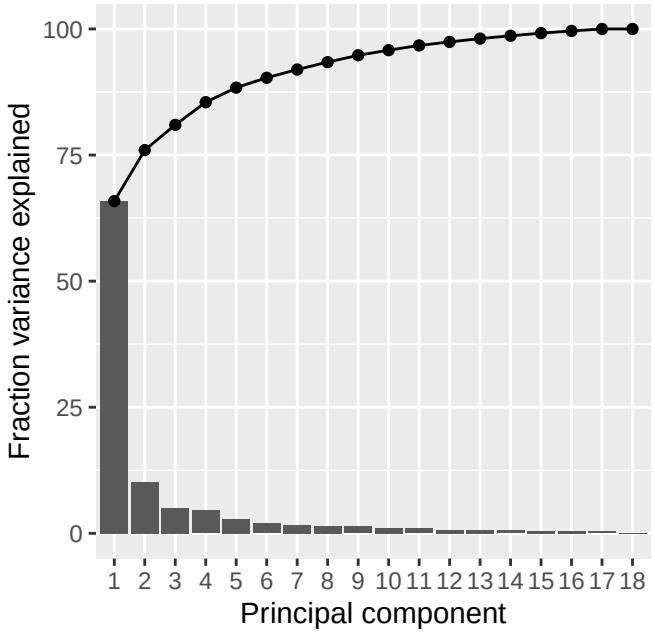
4.4.4.2 Chemokine related genes There is a lot of research targeting the role of chemokines during neuronal development. We discovered ENST00000337225, which encodes for CXCL14 (alpha class chemokin ligand). ENST00000337225 shows a significant increase in expression between weeks 4 to 6 and is related to Anterior cingulate cortex. ENST00000579298 encodes for NUP85 (nucleoporin 85), a protein component of the Nup107-160 subunit of the nuclear pore complex. NUP85 can bind to CCR2 (a receptor for beta class chemokines) and promotes chemotaxis of monocytes. ENST00000579298 is related to the Frontal Cortex and shows a decline between weeks 5 to 9.

4.5 Hypothesis: Specific TRAs can be used as key biomarkers for the development of individual organs

4.5.1 Principal component analysis

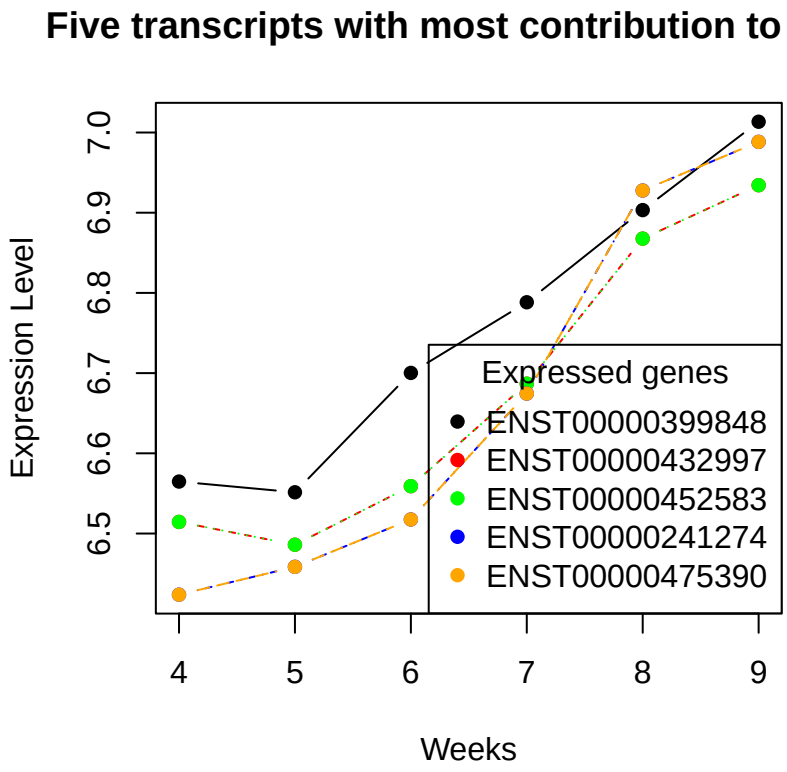
Principal component analysis (PCA) was performed on a matrix containing all differentially expressed genes throughout the 6 weeks. This was performed to reduce dimension while keeping most of the data's variance.

To get a better grasp of how much variance is explained by each PCA the percentage variance of each PCA (scree plot) together with the cumulative variance of the PCAs together was visualized in a plot.



This visualization helps to determine how many PCAs are needed to explain a significant amount of the data's variation. The chosen PCAs can than be used for further analysis. The first three PCAs already explain over 80% of the data's variance. Notably, the first PCA alone already explains over 60% of the data's variance and all following PCAs explain significantly less variance(Figure **??**).

For further analysis, each transcript was awarded a rank depending on how much it contributes to a specific PCA. We took a closer look at the five most contributing transcripts of PCA1, as this PCA alone already explains the brunt of the variance of our data. These five transcripts with the highest contribution encode for three different proteins: Low-Density Lipoprotein Receptor Class A Domain-Containing Protein 4 (LDLRAD4) encoded by ENST00000399848, Calcium-activated potassium channel subunit beta-2 (KCNMB2) encoded by ENST00000432997 and ENST00000452583 and SLIT And NTRK-Like Protein 3 (SLITRK3) encoded by ENST00000241274 and ENST00000475390.



The transcripts of ENST00000432997 and ENST00000452583 encode for the same Protein (KCNMB2). Fitting to that, both transcripts consistently display the same expression level over all weeks. The same applies to the transcripts of ENST00000241274 and ENST00000475390, as they also encode for the same protein (SLITRK3) and therefore also display the same expression levels. The transcripts for LDLRAD4 show the highest overall expression levels. Their expression levels also don't display many similarities to those of KCNMB2 and SLITRK3, as their increase patterns differ. However, KCNMB2 and SLITRK3 show great similarity in their expression levels over all weeks. Especially between weeks 5 to 7 KCNMB2 and SLITRK3 show near identical expression patterns.

5 Discussion

5.1 Comparison with original paper

In contrast to our *limma* analysis to generate data with biological relevance, the authors of the original paper (Yi H *et al.*, 2010) of our data set used *One-way analysis of variance* to acquire differential expressed genes. We compared both data subsets by k-means clustering and plotting the overlapping Intensity values (Figure 4). By looking at the trends of the cluster for *up* and *down* regulated gene expression, we can validate the method used by the authors. In contrast the *arch* regulated genes-clusters do not present a clear result and show a weakness of the method used. It falsely detects some genes as differentially expressed, assuming our *vsn* normalization is correct, which is indicated by the quality control (section 3.3.1).

Furthermore, we showed in our Venn diagram (Figure 3), that our method detected differentially expressed genes not included in their data. This questions the significance of either the papers results or again our *vsn* normalization and subsequent *limma* analysis, which is rather unlikely given our quality control and plots.

The discrepancy can be explained by the publication date, which is year 2010. Now there are more advanced algorithms that determine differentially expressed genes more precisely. The *vsn* and the *limma* package are up to date with frequent advances doi: 10.1093/nar/gkv007. Overall, this ensured the quality of our data set and differentially expressed genes while also pointing out some flaws of the data set of the paper.

5.2 Hypothesis: TRAs can infer a timeline of organ development similar to the results by Yi et al. 2010

In our analysis, we have shown that a number of TRAs are differentially expressed (section 4.3.1) between week 4 and 9 of human embryonic development in each of the analyzed tissues. Nonetheless, the expression levels of TRAs associated with one tissue do not constitute a useful metric for the organ's development (section **??**). This can be explained by the fact that within one tissue's TRAs, there are multiple groups of genes both distinct in expression patterns (clustering in section **??**) and function (analysis of spleen gene functions in section 4.3.3). Thus, we determined that the expression over time of functional gene sets linked to specific tissues through overrepresentation analysis is a more meaningful metric for organ development.

This approach was used in section 4.3.4 for eight different tissues. For the spleen, the results of our analysis (Fig. **??**A) largely do not reflect the embryonic development (section 2.2). While some of the immune-related gene sets are already expressed in week 4, the spleen only develops by week 6 and contains immune cells by week 12. This shows that while the spleen plays a role in the immune system and such gene sets are therefore rightly linked to the spleen, the expression of these transcripts alone does not necessarily relate to the development of the organ. It is still noteworthy that functions related to the adaptive immune system increase in expression form week 7 onward, which correlates with the beginning of T-cell development in the thymus. The observed timeframe is an important part of brain development (section 2.2). This is also visible in the expression data (Fig. **??**B), with a already high but still continously increasing expression of synaptic gene sets. Furthermore, as the brain starts to form, the expression of neuron projection morphogenesis transcripts increases continuously from week 5 to 8.

At week 4, the clearly heart-associated gene sets (Fig. **??**C) are at their highest expression level and decrease until week 8. The cardiac muscle tissue development transcripts still remain highly expressed (>7.5). This corresponds to the early development of the heart as noted in the introduction (section 2.2). It is noteworthy that the heart contraction gene set rises in expression again from week 8 to 9, but here an explanation is not possible without further analyzing the individual genes. The liver-associated TRAs showed no clear expression pattern (Fig. **??**D). Thus, even though the liver forms mostly during the analyzed timeframe (section 2.2), we cannot link the gene expression to the organ's development. The detected functions are mostly metabolic pathways whose activity could also be related to processes outside the liver. As a result, it is plausible that their expression is independent of liver development. The skeletal muscle functions are expressed only late within the observed time, as shown by the large increase in expression from week 8 to 9 (Fig. **??**E). As muscle fibers begin to develop later than week 9 and the first related proteins appear from week 7 on (section 2.2), these expression data correspond well to the embryonic development. The testis gene sets decrease in expression from week 5 onward (Fig. **??**E). This is in contrast to the embryonic development, where the gonads start to form at around the same time (section 2.2). For the stomach, the expression pattern indicates a decrease until week 6 followed by rising expression levels until week 9 (Fig. **??**G). However, the literature indicates that these results are unrelated to the stomach development. Functions like digestion or peptide hormone secretion are impossible to occure at this time, since the specific cells needed for this only appear later in embryogenesis (section 2.2). Therfore, the cause of the changing expression would have to be determined through a more in-depth analysis of the involved genes. Finally, the skin shows an increased expression of related genes sets from week 5 through 9 (Fig. **??**H). This broadly reflects the embryonic development, with the epidermis starting to form in week 4 (section 2.2). We also found this expression pattern in the keratinization gene set that is suggested by literature as a good indicator for skin formation.

5.3 Hypothesis: Neural TRA expression patterns reflect morphological brain development

First point to discuss is the significant increase in the gene expression of Ca(2+)ATPase (SLN sarcolipin) by factor 2 in logarithmic scale (Fig. 8 left). A cause of this might be the process of neuronal migration which starts at week 9 (section 2.2). Ca(2+) is an essential cofactor for actin dependent cell migration.

Another point is the strong correlation between the expression of SNAP91 genes and one SYBU gene (ENST00000276646) (Fig. 8 middle). Both proteins were associated with the Cerebellar Hemisphere and contribute to endocytosis, which is essential for functional neurons and contributes to neuronal survival (Overhoff et al. 2020). One point to mention here is, that ENST00000521485 and ENST00000518312 were both associated with SNAP91, nevertheless to different isoforms. But both show identical correlation in expression, hence a failure in annotation might be possible.

In addition we identified a significant increase in expression of NXPH1 (Fig. 8 middle), this refers to the NXPH1-promoted adhesion between axons and dendrites. An upregulation of this supplementary factor can prepare the process of synapse formation which starts at week 11 (section 2.2). We further identified a strong LSAMP expression associated to the putamen , which is associated to the putamen (Fig. 8 middle). The recognized increase in gene expression of LSAMP starting at week 6 might be a preparation for synapse formation at week 9, hence LSMP is an adhesion molecule in axon guidance (section 2.2).

We further identified that two transcripts associated to two different tissues show identical correlation, this can be caused by false annotations. Nevertheless, they show a strong downregulation of BDNF (Fig. 8 middle). This factor normally promotes neuronal survival. A downregulation of BDNF can be a preparation for the phase of programmed cell death, which starts at week 20 in neuronal development (section 2.2).

The neuronal cell migration is strongly dependent on chemoatractors like chemokines (Tiveron 2008). We identified a significant increase in CXCL14 between weeks 5 to 6 and a maintaining high expression level for the following weeks (Fig. 8 right). This could be an accumulation for neuronal migration, starting at week 9 (section 2.2). A decline in NUP85 expression is noteable between week 5 to 9 (Fig. 8 right). NUP85 can bind to CCR2, hence a decline in NUP85 genexpression reduce the chances for this binding. A consequence might be that more CCR2 receptors are free for beta type chemokin mediated signals.

5.4 Hypothesis: Specific TRAs can be used as key biomarkers for the development of individual organs

6 References

Benninghoff, A. (1993), Makroskopische Anatomie, Embryologie und Histologie des Menschen, 15. Auflage, München, Wien, Baltimore, Urban und Schwarzenberg.

Deutsch J. (2013), Embryologie und Physiologie der Leber, Pädiatrische Gastroenterologie, Hepatologie und Ernährung, 375–87, Berlin Deutschland.

Dinkelacker M., (2019), Chromosomal clustering of tissue restricted antigens.

Esrefoglu M., Taslindere E. & Cetin A. (2017), Development of the Esophagus and Stomach, Bezmialem sci. 5, 175-82.

Fasold M. & Binder H., (2013), AffyRNADegradation: control and correction of RNA quality effects in GeneChip expression data, Bioinformatics, 29, 129-31. Hayward AR., (1983), The human fetus and newborn: development of the immune response, Birth Defects Orig. Artic. Ser. 19, 289-94.

Hikspoors J.P.J.M., Kruepunga N., Mommen G.M.C. et al., (2022), A pictorial account of the human embryonic heart between 3.5 and 8 weeks of development, Commun. Biol. 5, 226.

Hu MS., Borrelli MR., Hong WX., Malhotra S., Cheung ATM., Ransom RC., Rennert RC., Morrison SD., Lorenz HP. & Longaker MT., (2018) Embryonic skin development and repair, Organogenesis 14, 46-63.

Huber W., von Heydebreck A., Sültmann H., Poustka A., Vingron M., (2002), Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics. 18 ,Suppl 1, 96-104.

James F. & Jones M.D., (1983), Development of the Spleen, Lymphology 16, 83-89, Georg Thieme Verlag Stuttgart, New York.

Kluth D., Jaeschke-Melli S. & Fiegel H., (2003),The embryology of gut rotation, Semin. Pediatr. Surg. 12, 275-279.

Müller W. A. & Hassel M., (2018), Entwicklungsbiologie und Reproduktionsbiologie des Menschen und bedeutender Modellorganismen, 6. Auflage, Springer-Verlag GmbH Deutschland, Berlin, Deutschland.

Murphy K. & Weaver C., (2018), Janeway Immunologie, 9. Auflage, Springer-Verlag GmbH Deutschland, Berlin, Deutschland.

National Research Council and Institute of Medicine, (2009), Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities, Washington, DC: The National Academies Press.

Overhoff M., De Bruyckere E. & Kononenko N- L., (2020), Mechanisms of neuronal survival safeguarded by endocytosis and autophagy, J. Neurochem. 157, 263-296.

Ritchie ME., Phipson B., Wu D., Hu Y., Law CW., Shi W. & Smyth GK., (2015), limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43, e47.

Romero N. B., Mezmezian M .& Fidziańska A. (2013), Pediatric Neurology Part III, Chapter 137 - Main steps of skeletal muscle development in the human: Morphological analysis and ultrastructural characteristics of developing human muscle, Handb. Clin. Neurol. 113, 1299-1310.

Tiveron M.C. & Cremer H., (2008), CXCL12/CXCR4 signalling in neuronal cell migration, Curr. Opin. Neurobiol. 18, 237-244.

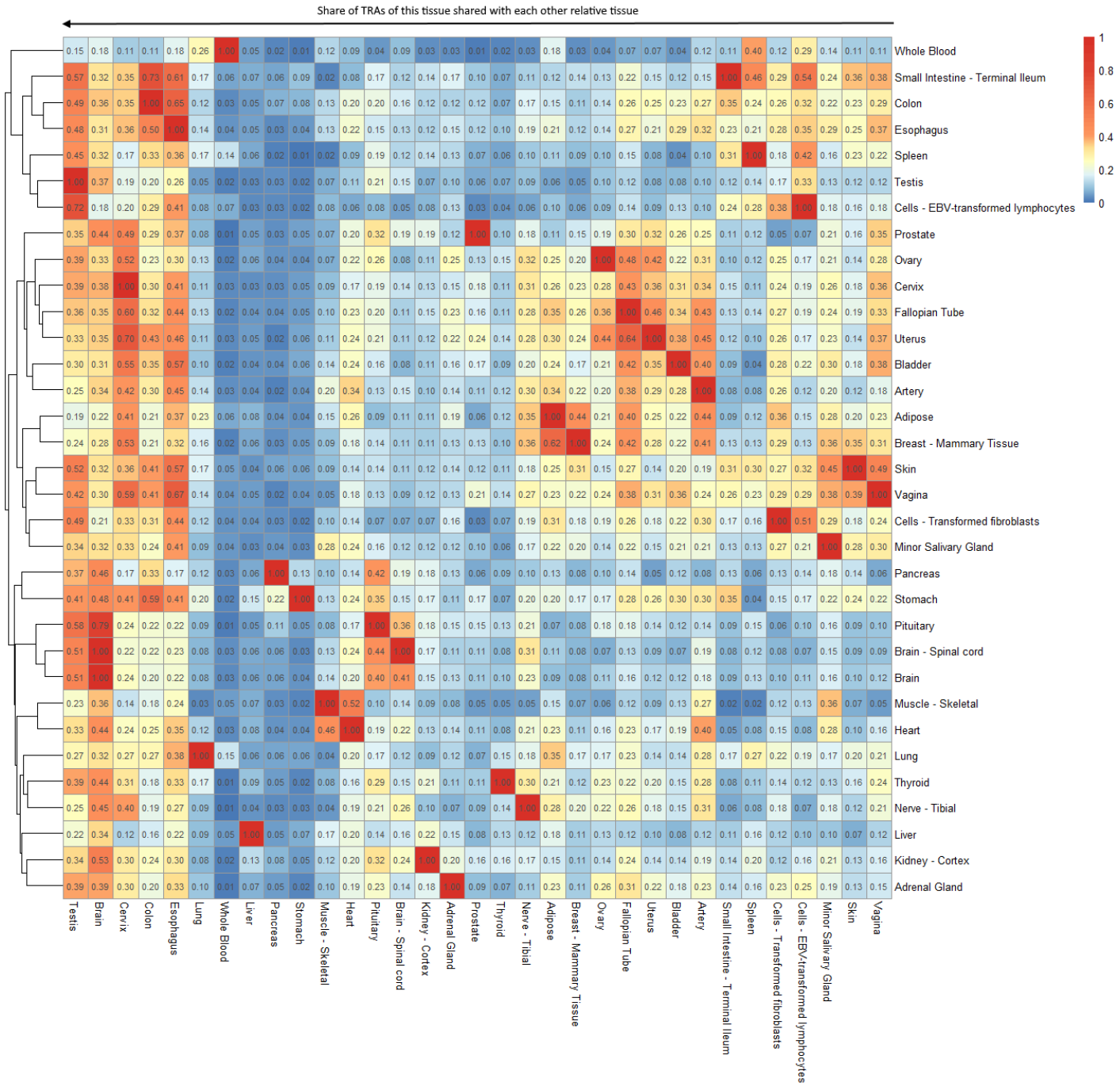
Ulfig N., (2009), Kurzlehrbuch Embryologie, Georg Thieme Verlag Stuttgart,New York.

Yi H., Xue L., Guo MX., Ma J., Zeng Y., Wang W., Cai JY., Hu HM., Shu HB., Shi YB. et al., (2010), Gene expression atlas for human embryogenesis, FASEB. J. 24, 3341-50..

Yusuf F., Rehimi R., Dai F. & Brand-Saberi B., (2005), Expression of chemokine receptor CXCR4 during chick embryo development, Anat. Embryol. 210, 35-41.

7 Supplementary

7.1 Organ development



7.2 Brain development

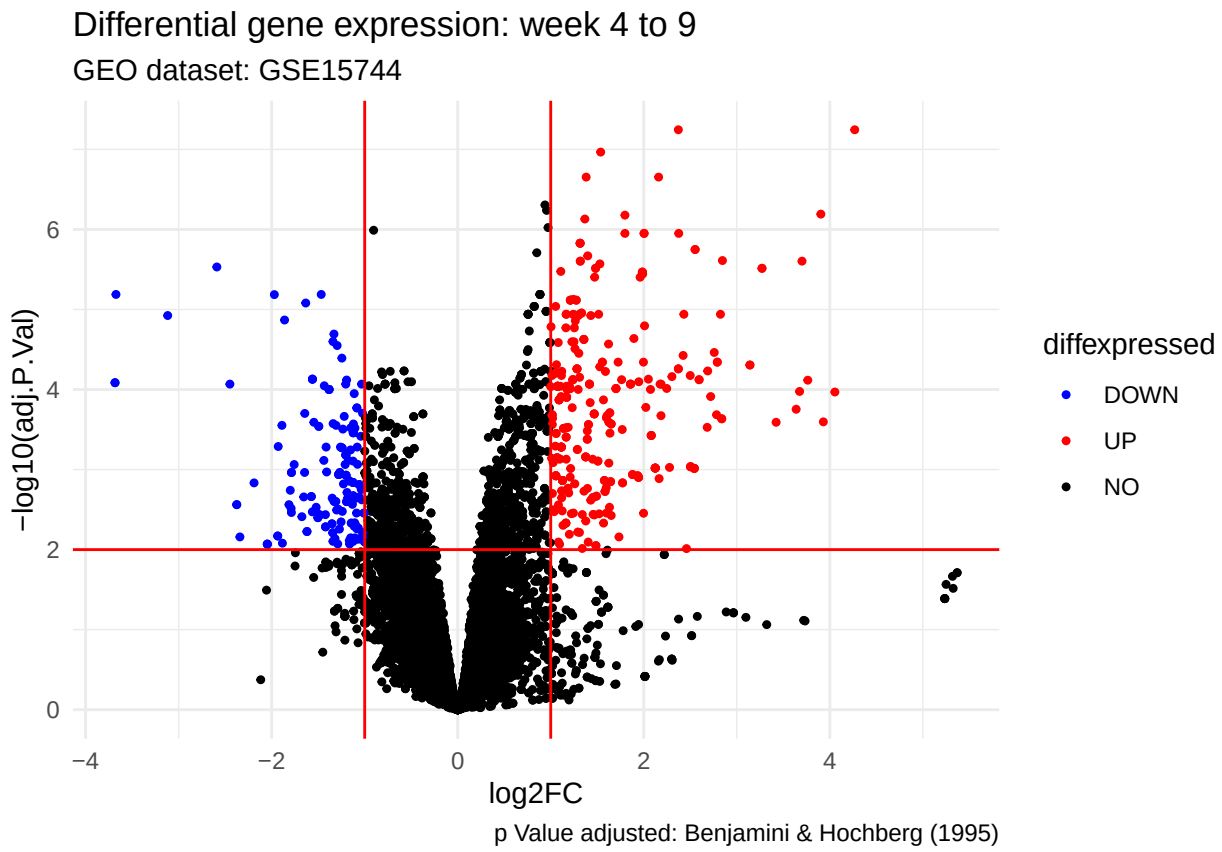
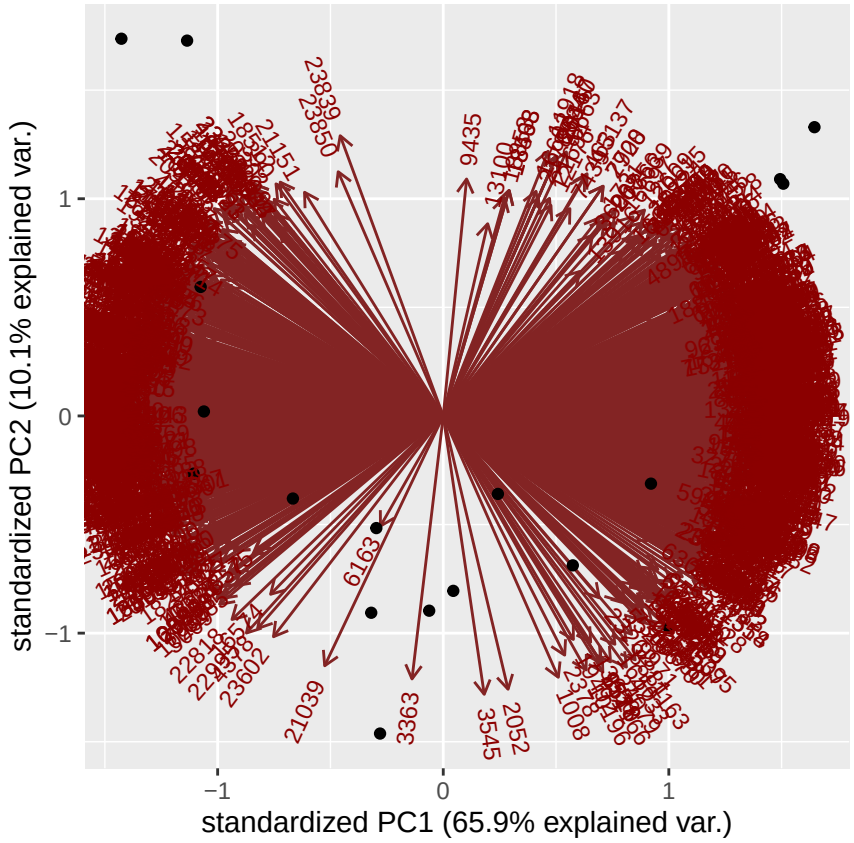
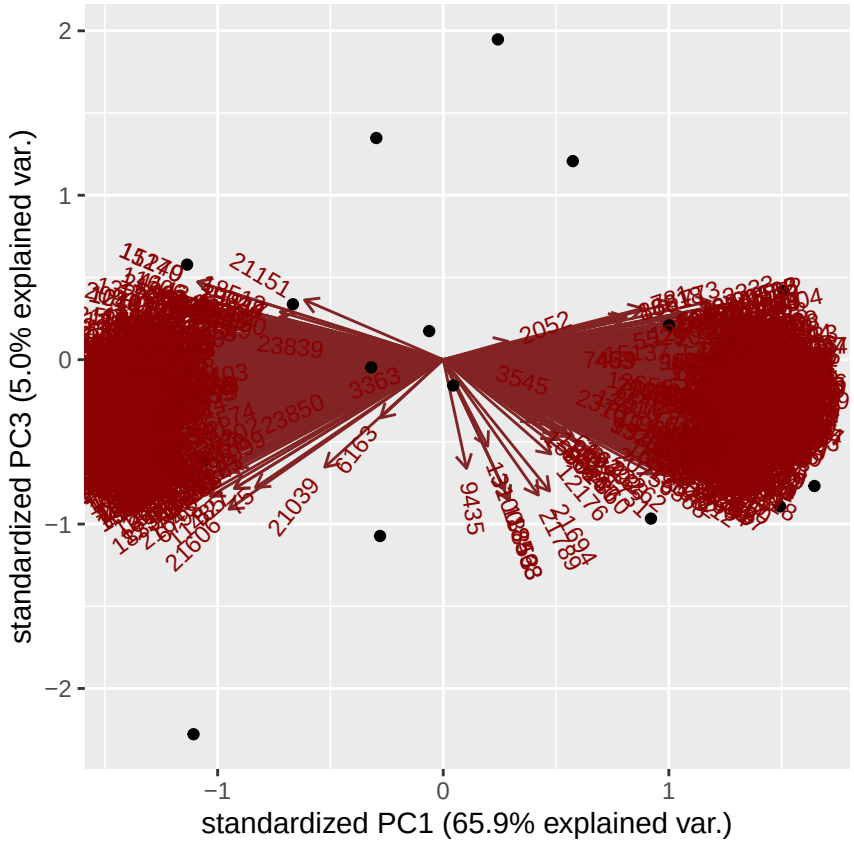


Figure 9: *Volcanoplot for differential gene expression between week 4 and 9. The adjusted P-value boundary was set at 0.01. The logFC boundarys were -1 and 1.*

7.3 PCA

The first three PCAs were now plotted against each other. This helps to analyze the genes contributions to the PCAs. The more parallel a gene's vector is to the PCA axis, the more the gene contributes to that PCA. The lenght of the vectors shows how much variability of the gene is explained by the PCAs. The longer the vector, the better it is represented in this dimension. The angels between the vectors give insight about the correlation between the genes. Small angels demonstrate a high correlation whereas opposite angles demonstrate a high negative correlation.


standardized PC2 (10.1% explained var.)
standardized PC1 (65.9% explained var.)
23839
23850
9435
21151
6163
21039
3363
3545
2052
23602
22818
1008


standardized PC3 (5.0% explained var.)
standardized PC1 (65.9% explained var.)
21151
23839
3363
2052
3545
23850
6163
9435
21039
12176
21694
21789
21606