Test3

Paul Christmann

2022-07-12

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.

Normalising the data set

Intensity values of different chips are affected by sample preparation and array manufacturing and processing resulting in statistical variance and random fluctuation. To access the biological relevant variation the raw data needs to be transformed by normalisation. We chose the vsn rma normalization with its library *vsn* according to Huber *et al.* (2002).

The library vsn 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.

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*.

Limma package

The *limma* package is installed and imported using bioconductor. Among many other things, it determines changes of gene expression over time in intensity values of microarrays, so called differential expressed genes. It facilitates advanced statistical algorithms to calculate the necessary coefficients of a linear model for every intensity value in the data set. It does this by utilizing information borrowing, quantitative weighting, variance modelling and data preprocessing, but importantly, it does not subset the data (Ritchie. et al. 2015) Because the same linear model was performed on every intensity value, statistical tests called Empirical Bayes can determine differential expressed genes via t-statistics and their associated p-values, while reducing the variance of residuals.

Over representation analysis

The statistical method over-representation analysis

Results

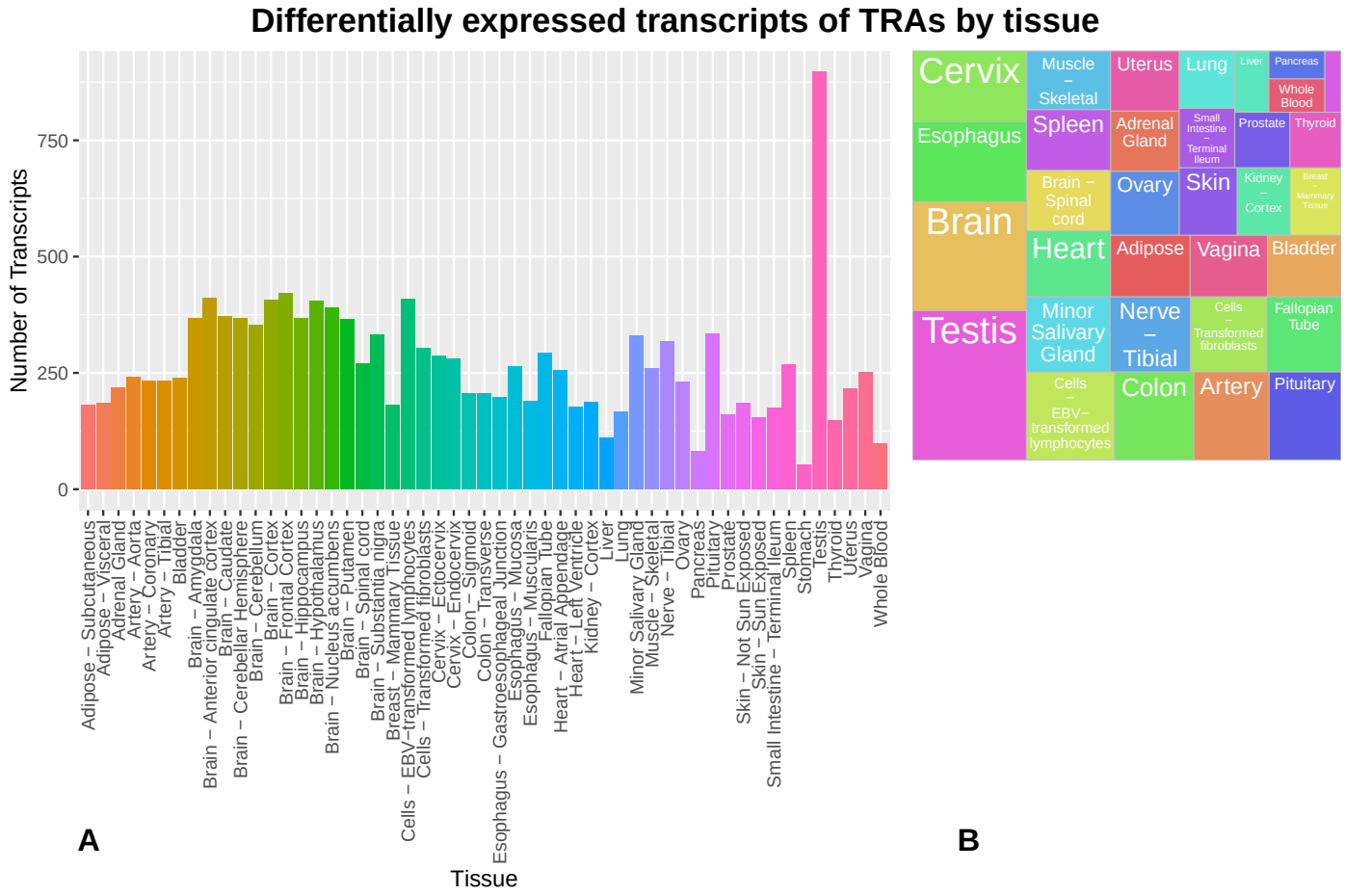
Limma analysis

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.

TRAs can infer a basic timeline of organ development

Differentially expressed transcripts can be linked to all analyzed tissues

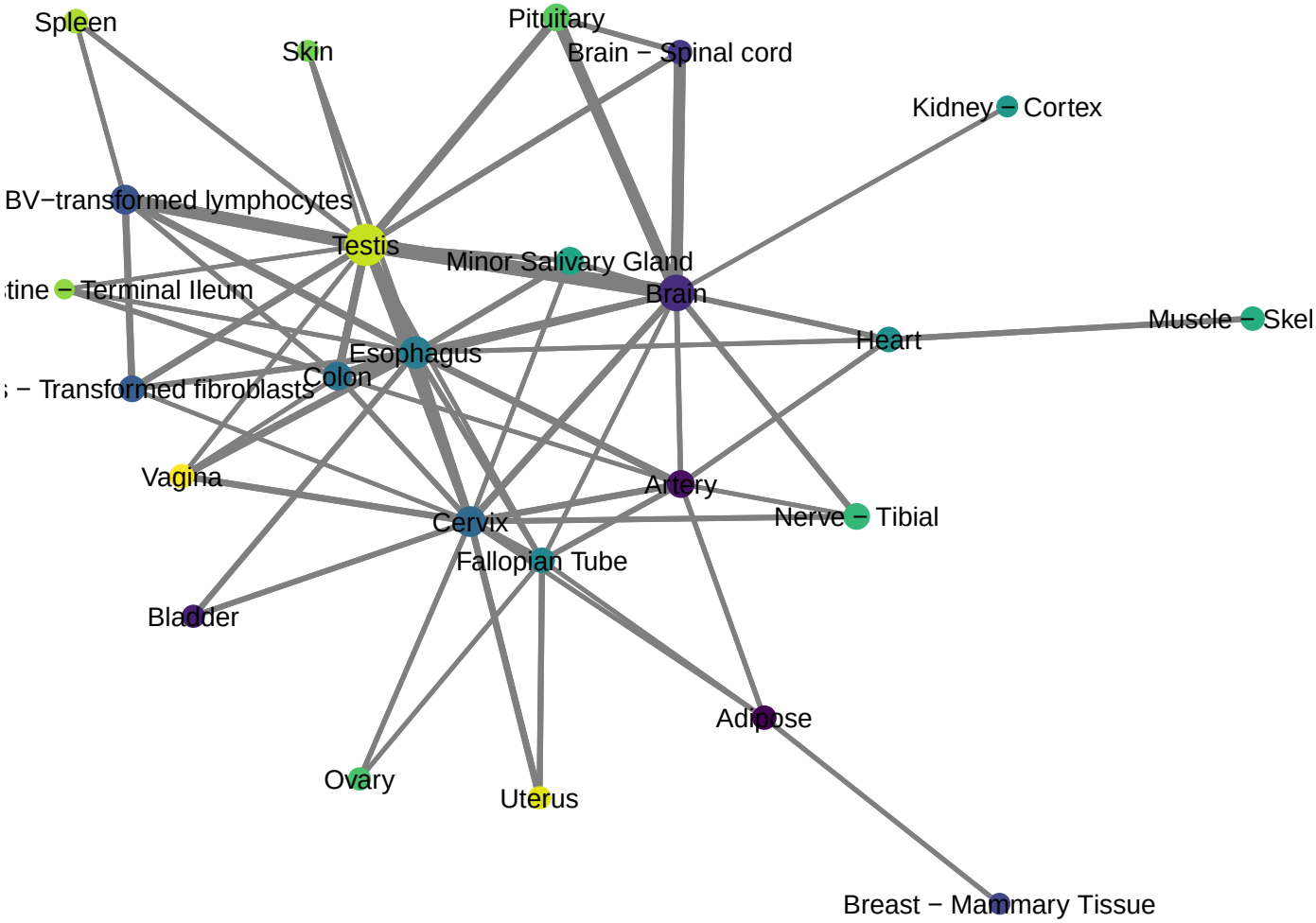
The TRA Data covers 53 distinct tissues. For all of those, we found at least 40 differentially expressed transcripts within our dataset. The minimum was found with 46 stomach-linked transcripts, the maximum were 837 TRAs for the testes.




```
[1] 7.356921
```

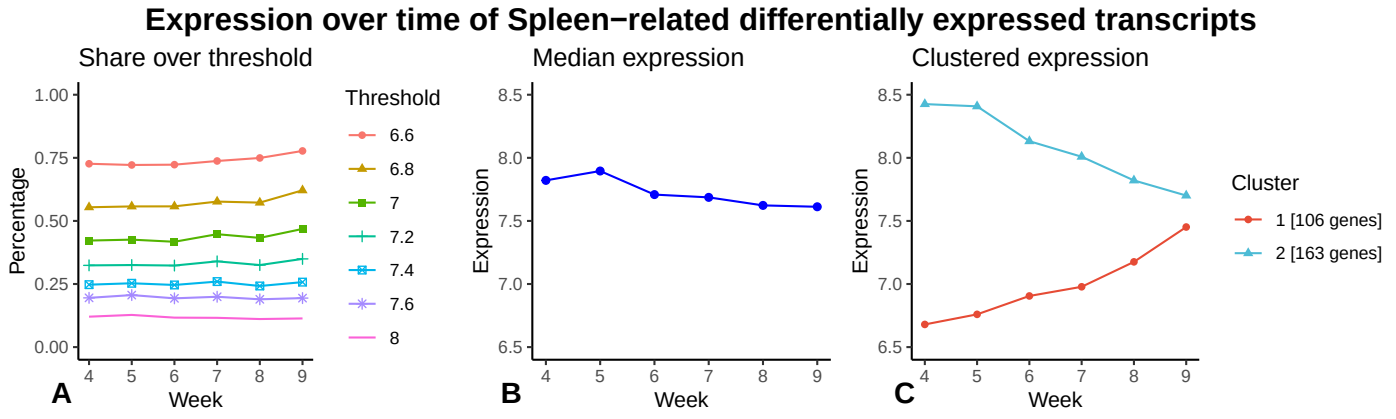
These numbers are sufficient for further analysis of the gene expression within individual tissues. Therefore, all further analysis will be based on our dataset with differentially expressed genes from limma analysis. Nonetheless, it should be noted that there is a significant overlap between the TRAs 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. @ref(fig:tissue-links).

Overlap between the TRAs associated with different tissues



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.

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.

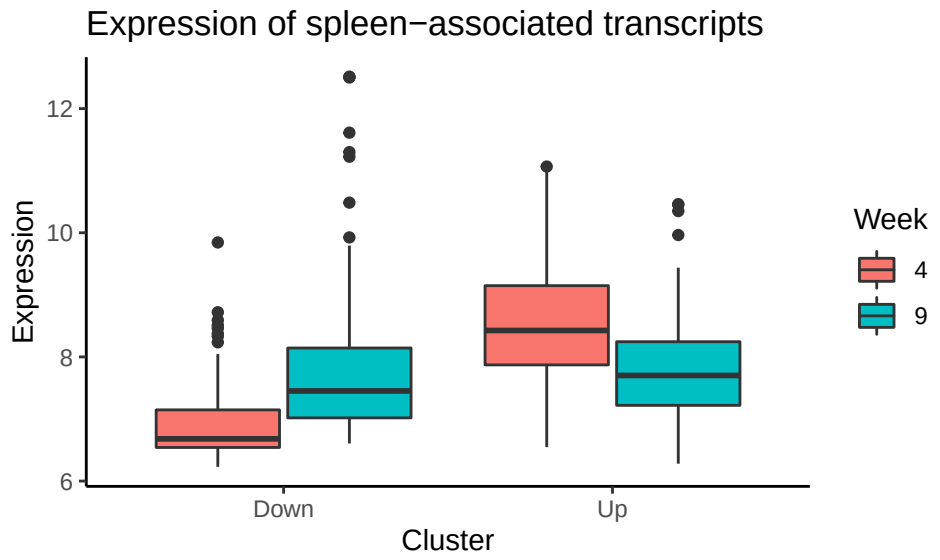


Figure 1: *A further look on the expression of transcripts in the up- and downregulated clusters shows that the upregulated transcripts are close to the minimum expression level between 6 and 7 in week 4 and showing expressions between 7 and 8.5 by week 9. In contrast, the downregulated genes have very 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. @ref(fig:spleen-boxplot), 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.].

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.

Discussion

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

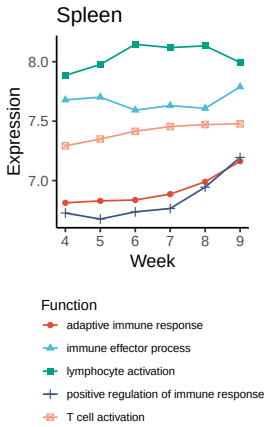
In our analysis, we have shown that a number of TRAs are differentially expressed (section @ref(organ-overview)) between week 4 and 9 of human embryonic development in each of the analyzed tissues. Nonetheless, the expression levels of TRAs associated with one tissue do not constitute a useful metric for the organ's development (section @ref(organ-clustering)). 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 @ref(organ-clustering)) and function (analysis of spleen gene functions in section

Main functions from overrepresentation analysis plotted by tissue

Spleen

FUNCTION	ADJ_P_VALUE
lymphocyte activation	1×10^{-65}
T cell activation	3×10^{-56}
adaptive immune response	4×10^{-53}
immune effector process	5×10^{-50}
positive regulation of immune response	8×10^{-50}
regulation of cell activation	3×10^{-49}
regulation of leukocyte activation	4×10^{-49}
regulation of lymphocyte activation	8×10^{-48}
leukocyte cell-cell adhesion	2×10^{-42}
immune response-regulating signaling pathway	1×10^{-41}

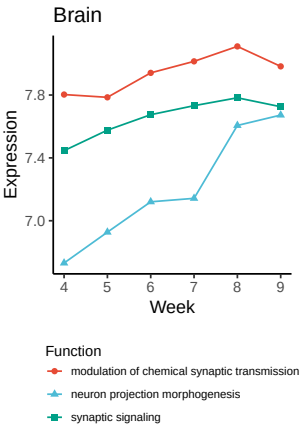
A



Brain

FUNCTION	ADJ_P_VALUE
synaptic signaling	9×10^{-43}
trans-synaptic signaling	9×10^{-43}
chemical synaptic transmission	9×10^{-43}
anterograde trans-synaptic signaling	9×10^{-43}
modulation of chemical synaptic transmission	9×10^{-25}
regulation of trans-synaptic signaling	9×10^{-25}
neuron projection morphogenesis	2×10^{-18}
cell projection morphogenesis	8×10^{-18}
plasma membrane bounded cell projection morphogenesis	2×10^{-17}
cell part morphogenesis	2×10^{-17}

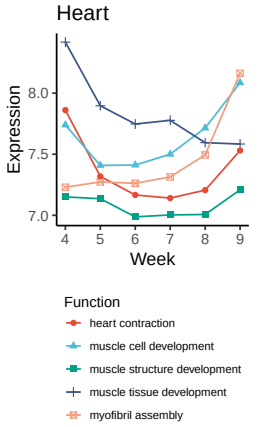
B



Heart

FUNCTION	ADJ_P_VALUE
muscle cell development	3×10^{-27}
muscle structure development	1×10^{-26}
myofibril assembly	2×10^{-24}
striated muscle cell development	2×10^{-24}
muscle tissue development	3×10^{-24}
cardiac muscle tissue development	3×10^{-24}
muscle system process	3×10^{-24}
striated muscle tissue development	2×10^{-23}
muscle cell differentiation	5×10^{-23}
heart contraction	1×10^{-22}

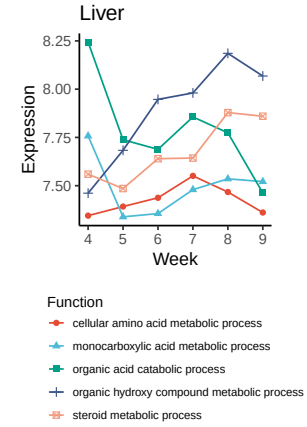
C



Liver

FUNCTION	ADJ_P_VALUE
monocarboxylic acid metabolic process	7×10^{-51}
small molecule catabolic process	7×10^{-39}
organic acid catabolic process	4×10^{-36}
carboxylic acid catabolic process	2×10^{-35}
steroid metabolic process	7×10^{-33}
small molecule biosynthetic process	4×10^{-31}
cellular amino acid metabolic process	1×10^{-28}
fatty acid metabolic process	6×10^{-28}
organic hydroxy compound metabolic process	9×10^{-28}
alpha-amino acid metabolic process	3×10^{-27}

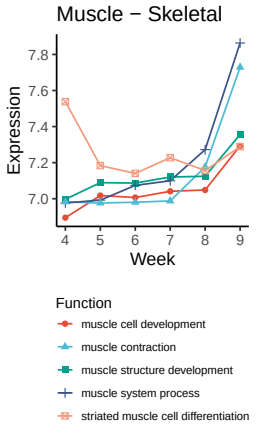
D



Muscle - Skeletal

FUNCTION	ADJ_P_VALUE
muscle structure development	7×10^{-29}
muscle system process	8×10^{-27}
muscle organ development	5×10^{-25}
muscle cell development	1×10^{-21}
muscle contraction	1×10^{-21}
muscle cell differentiation	4×10^{-20}
striated muscle cell differentiation	4×10^{-19}
myofibril assembly	3×10^{-18}
striated muscle cell development	3×10^{-18}
striated muscle contraction	1×10^{-16}

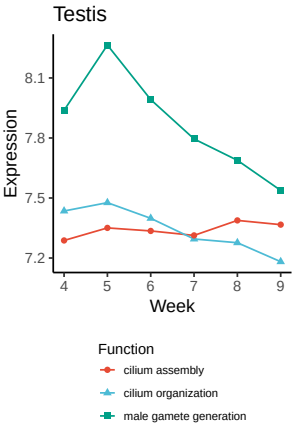
E



Testis

FUNCTION	ADJ_P_VALUE
male gamete generation	1×10^{-35}
spermatogenesis	3×10^{-34}
microtubule-based movement	2×10^{-26}
cilium organization	2×10^{-26}
cilium assembly	2×10^{-23}
microtubule cytoskeleton organization	2×10^{-23}
meiotic cell cycle	2×10^{-23}
cilium movement	3×10^{-20}
nuclear division	1×10^{-18}
meiotic cell cycle process	2×10^{-18}

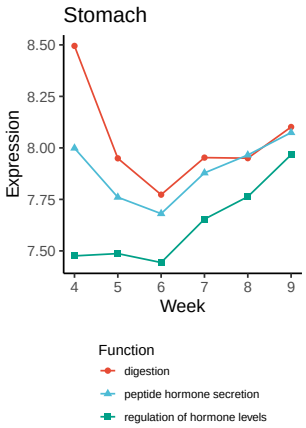
F



Stomach

FUNCTION	ADJ_P_VALUE
digestion	1×10^{-4}
regulation of hormone levels	6×10^{-4}
gastric acid secretion	2×10^{-3}
peptide hormone secretion	2×10^{-3}
peptide transport	2×10^{-3}
peptide secretion	3×10^{-3}
digestive system process	6×10^{-3}
amide transport	6×10^{-3}
hormone transport	6×10^{-3}
hormone secretion	6×10^{-3}

G



Skin

FUNCTION	ADJ_P_VALUE
epidermis development	2×10^{-21}
skin development	9×10^{-21}
keratinization	6×10^{-17}
keratinocyte differentiation	2×10^{-14}
epidermal cell differentiation	1×10^{-10}
intermediate filament cytoskeleton organization	9×10^{-10}
intermediate filament-based process	1×10^{-9}
intermediate filament organization	6×10^{-9}
regulation of water loss via skin	2×10^{-8}
establishment of skin barrier	1×10^{-7}

H

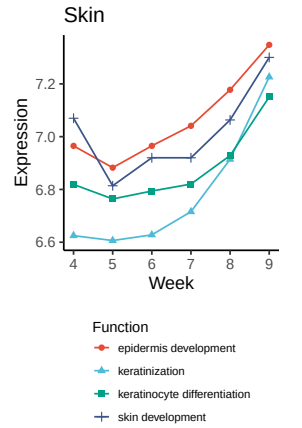


Figure 2: *For eight tissues A to H, the tables show the 10 most significant results from overrepresentation analysis. There, the functional GO annotation of all transcripts in our dataset (not only the differentially expressed ones) was compared to the functions associated with all TRAs of one tissue. The significance was determined by an adjusted p-value. Of those 10 functions, up to 5 were*

@ref(organ-tables)). 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 @ref(organ-ora) for eight different tissues. For the spleen, the results of our analysis (Fig. @ref(fig:ORA-plot)A) largely do not reflect the embryonic development (section @ref(intro-tissues)). 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 @ref(intro-tissues)). This is also visible in the expression data (Fig. @ref(fig:ORA-plot)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. @ref(fig:ORA-plot)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 @ref(intro-tissues)). 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. @ref(fig:ORA-plot)D). Thus, even though the liver forms mostly during the analyzed timeframe (section @ref(intro-tissues)), 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. @ref(fig:ORA-plot)E). As muscle fibers begin to develop later than week 9 and the first related proteins appear from week 7 on (section @ref(intro-tissues)), these expression data correspond well to the embryonic development. The testis gene sets decrease in expression from week 5 onward (Fig. @ref(fig:ORA-plot)E). This is in contrast to the embryonic development, where the gonads start to form at around the same time (section @ref(intro-tissues)). For the stomach, the expression pattern indicates a decrease until week 6 followed by rising expression levels until week 9 (Fig. @ref(fig:ORA-plot)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 @ref(intro-tissues)). 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. @ref(fig:ORA-plot)H). This broadly reflects the embryonic development, with the epidermis starting to form in week 4 (section @ref(intro-tissues)). We also found this expression pattern in the keratinization gene set that is suggested by literature as a good indicator for skin formation.