Test4

Paul Christmann

2022-07-12

Abstract

A single cell turns into a complex multicellular organism during embryogenesis. The morphological steps behind this process are mostly understood. The role of molecular biological pathways and molecules like chemokines in the embryonic development is quite unknown and has to be investigated. Tissue restricted antigens (TRA) may be a key in understanding the relation between molecular biological cause and morphological consequence.TRAs originate from negative selection in thymus, where medullary epithelial cells (mTECs) express them on their surface to remove self-antigen reactive T-cells. Those TRA represent a wide range of self-antigen, hence we can use them to determine the gene expression in our dataset (Dinkelacker 2019; Murphy & Weaver 2018).
First of all, we want to compare our methods, e.g. limma analysis to the methods of the paper, e.g. **One-way analysis of variance** to find differentially expressed genes. We want validate, if TRAs represent the timeline in the early stage of embryonic development between week 4 to 9. And give us an insight in the molecular biological process in those weeks. Furthermore, we want to identify, if the expression of specific gens can be related to milestones in neurogenesis. Last but not least, we want to illustrate the role of specific TRAs as biomarkers for organogenesis.

Introduction

TRAs as a tool to gain insight into embryonic development

Tissues restricted 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.

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.

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).

Methods

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		

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.

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.

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*.

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 @ref(fig:QC-surface-images-and-RNA-Degradation-Plot)A 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.

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 @ref(fig:QC-surface-images-and-RNA-Degradation-Plot)B).

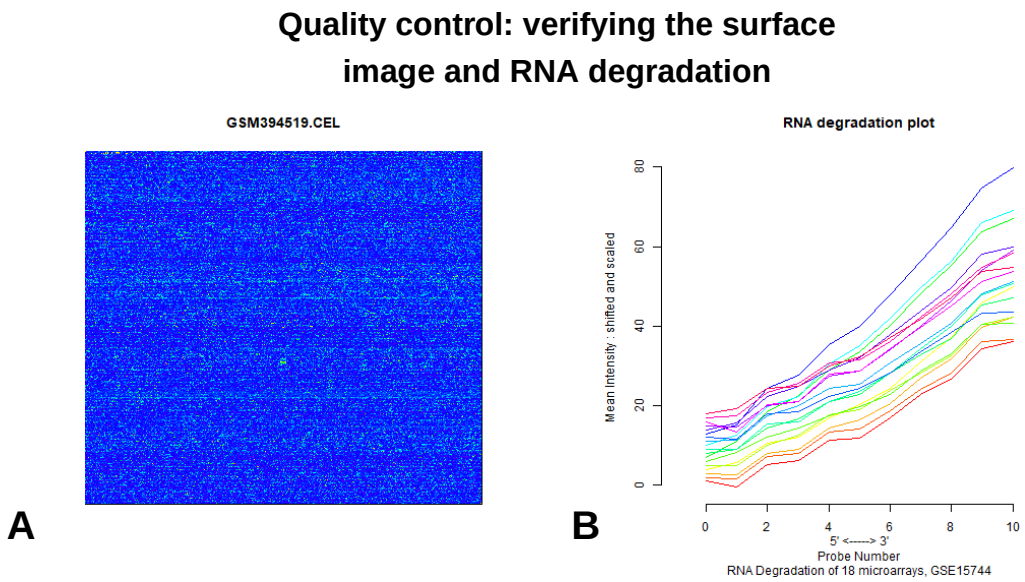

Figure 1: Quality control: Selected surface image of a microarray shows no damage or artefacts and RNA degradation plot shows slight irregularities and verifies the data. ***A****: The microarray inspection shows no irregularities and every chip is accepted for further data analysis.* ***B****: Some 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.*

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.

Quality control of the vsn normalization

To verify the transformed data intensity values, some test can be performed (Figure @ref(fig:QC-normalization-plots)). 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 @ref(fig:QC-normalization-plots)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 @ref(fig:QC-normalization-plots)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 @ref(fig:QC-normalization-plots)C).

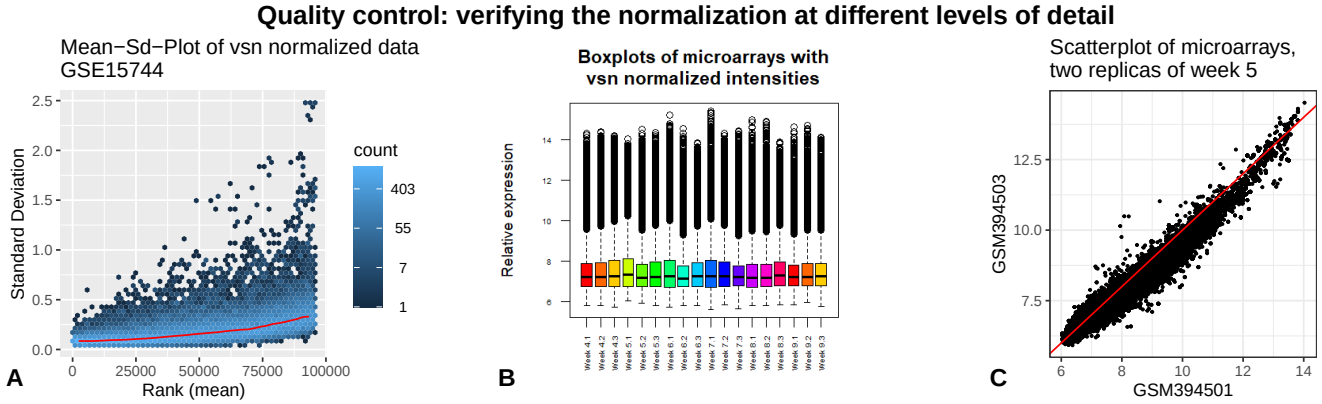

Figure 2: *The plots support the use of the vsn normalization on our data set. A: The red line is close to horizontal, although it shows some correlation at high intensity leves. B: The boxplots show nice allignment of the mean intensity values. Some outliers are given but can neglected given the 0.25 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*

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*.

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.

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.

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.

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 @ref(fig:Volcano-Plot)). 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.

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.

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 @ref(fig:Venn-Diagram)).

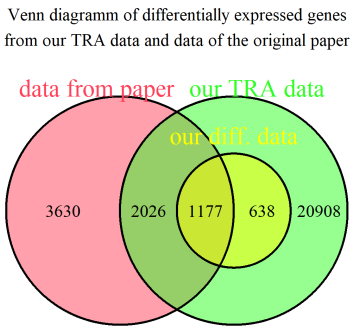

Figure 3: The Venn diagram shows overlapping sections between our data and the data from the original paper. *According to the paper, of the its 6,833 transcripts are differentially expressed. This overlapps with our general data with all TRAs (24,749 transcripts) and with our differentially expressed (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.

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 (Yi H *et. al*, 2010). 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 @ref(fig:paper-trends-cluster)).

We repeated the Figure@ref(fig:paper-trends-cluster)A 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 @ref(fig:paper-trends-cluster)C 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 @ref(fig:paper-trends-cluster)D), we recieve only 10% of the amount of transcripts which are clearly clusters.

Figure @ref(fig:paper-trends-cluster)E 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.

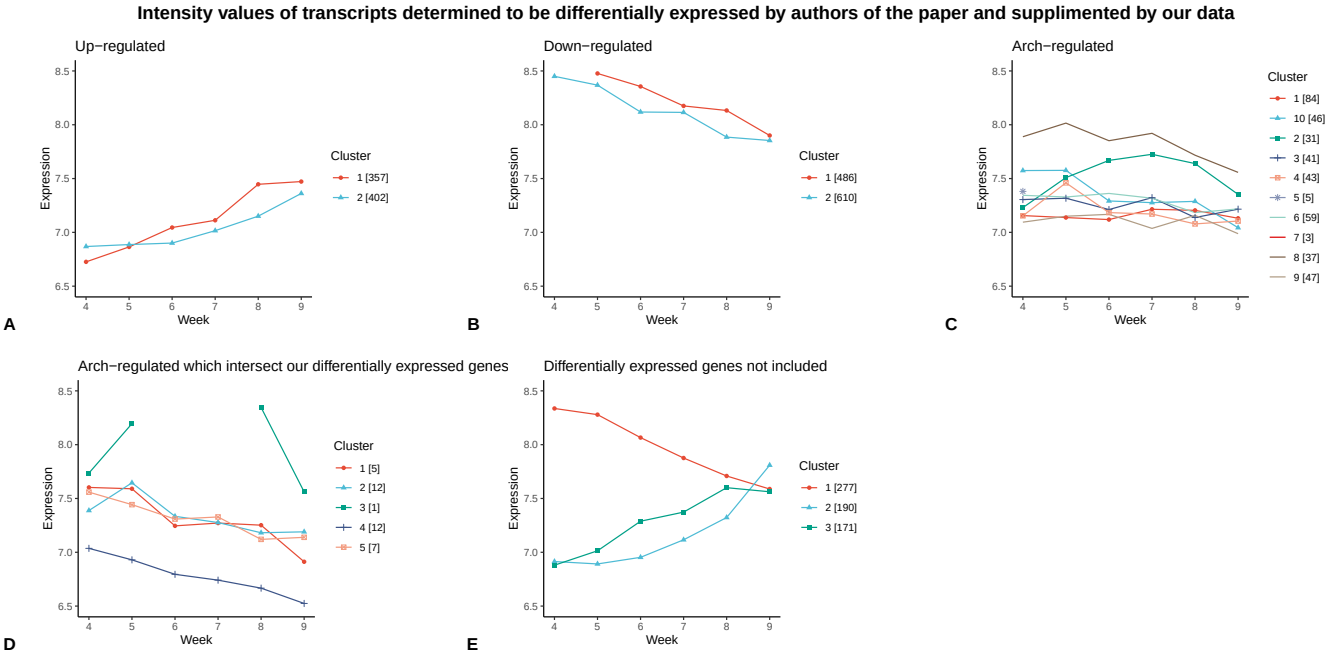

Figure 4: Clustering of the common transcripts reveals similar trends and but our additional data shows missing differentially expressed genes. *We chose data with transcript IDs both in the data set of the original paper and in our TRA data. Additionally they had to be annotated as* up, *down*, or *arch* 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*