Tissue Restricted Antigens as a tool in embryo research

Alina Aksianova, Lydia Steiner

Summer Semester 2022

Tutor: Ian Dirk Fichtner

Supervisor: Frau Dr. Maria Dinkelacker

Abstract

Embryogenesis is difficult to study and therefore still not fully understood. Yet, a novel approach borrowed from immune biology may provide the necessary tools to embryo research in the scope of an interdisciplinary application. Tissue restricted antigens are utilised by the immune system for the production of t-cells to prevent autoreactive events. Therefore, they are expressed simultaneously with other genes in the corresponding tissue. Tissue specific antigens seem to be a promising biological indicator to depict the intensity of gene expression to a given time during tissue development. Such markers would be a useful method to not only further analyse and understand embryogenesis and organogenesis but inspect general tissue development with high accuracy.

This project aims to explore embryogenesis in mice by analysing the expression levels of tissue restricted antigens over time and summarising these findings in an applicable dataframe.

1 Introduction

The immune system is responsible for recognising and eliminating threatening pathogens, such as bacteria, viruses or fungi, however, distinguishing these from endogenous cells and not harming them. CD8+ T cells (CD8s), also referred to as cytotoxic T cells precisely recognise antigens via specific T cell receptors (TCRs). Thus, playing a vital role in this process. In order to prevent an autoreactive self-antigen-CD8 complex that would be harmful to the organism, T cells undergo negative selection in addition to positive selection, in the thymus. This negative selection process is mediated by medullary thymic epithelial cells that are presented to developing T cells, leading to the elimination of binding T cells (mTECs)[@Kyewski]. Auto-antigens fall in two categories: housekeeping antigens (HAs) and tissue restricted antigens (TRAs). While HAs are expressed in a multitude of tissues, TRAs are found to be expressed rather uniquely. This results in HAs experiencing only insignificant amounts of epigenetic inactivation. TRAs in contrast, display silencing extensively in a majority of tissues. To classify as a TRA, an antigen is required to exceed five times the median gene expression within one to four tissue expression profiles [@dinkelacker2019]. An unpublished dataset by Dr. Maria Dinkelacker will be utilised to analyse embryonic development. Antigen expression levels constantly vary during embryogenesis, since different stem cells undergo differentiation. This includes TRA expression levels that display a temporal connection to the development of specific organs. QUELLE Due to this connection they offer an interesting approach to study organ development. Despite spatiotemporally mapping of organogenesis in mice recently, specific timing and expression levels of TRAs remain underexplored. Hence, this project aims to investigate TRA expression profiles within mice over a course of mid- to late embryogenesis. Data by @irie2011 supported this undertaking. This project aims to utilise the datasets on time-dependent embryonic transcriptomes and TRAs to catalogue established TRAs in respect to specific stages of embryonic development to offer a base of data that may support further studies in tissue development.

2 Methods

Datasets

Work by @irie2011comparative provided us with data to study the expression-levels of tissue-restricted antigens (TRAs) during embryonic development. TRA data was isolated from the Microarray data with the help of a TRA dataset [@dinkelacker2019]. The statistical open source programming language R [@R], bioconducter packages, and an annotation file [@cunningham2022ensembl] served to prepare data for the following analysis.

embryonic data set GSE28389 There were two main criteria for choosing a dataset. Firstly, it had to use Affymetrix microarray analysis. Secondly it was required to display expression levels of the whole mouse during different stages of embryogenesis. For the purpose of this work the whole RNA of multiple wild type C57BL/6 mice embryos was collected at eight different stages (Microarray chips: 3x E7.5, 3x E8.5, 3x E9.5, 3x E10.5, 2x E12.5, 2x E14.5, 2x E16.5, 2x E18.5). Triplicates or Duplicates were homogenised before application on the Affymetrix Mouse Genome 430 2.0 Array [@irie2011comparative].

TRA dataset Our TRA data stems from unpublished data by Dr. Maria Dinkelacker. The data included in the "tra.2014.mouse.4301.5x.table.tsv" table includes a larger quantity of TRAs and was thus chosen to work with.

Packages

The essential Bioconductor packages for the analyses were "affy" (version 1.74.0), "limma" (version 3.52.1) and "vsn" (version 3.64.0). The package "affy" was applied for exploratory oligonucleotide array analysis [@gautier2004irizarry]. "limma" was used for differential expression analysis of microarray data. The methods provided by the package give stable results even when the number of arrays is small like in this project [@ritchie2015limma]. "Limma" can be utilised for all gene expression technologies including microarray data. Variance stabilising normalisation via package "vsn" was employed for data normalisation. VSN executes data calibration, quantification of differential expression, and the quantification of measurement error [@huber2002variance].

Quality control (QC) of the embryonic microarray data

Four major quality complications are commonly encountered when working with microarray data. Low quality chips, imprints such as fingerprints or marks from pipette tips, irregularities in dye distribution, and light intensity extremes. To detect possible damages, we performed single-chip control for all twenty chips. Conspicuous chips should be removed at this point, however the number of biological replicates contained in the data was quite scarce. Some development stages only included data on two chips. Excluding an entire chip could thereby affect the significance of further statistical work. Single Chip Control There are no visual damages such as fingerprints present on the chips. Nevertheless the Affy chips of E12.5_1 and E14.5_1 are considerably lighter in intensity. This is also observable in the boxplot before normalisation.

RNA Degradation

RNA Degradation plots serve as another possibility to detect physical artefacts on the chips. It stands out that probe intensities are lower towards the 5' end of a transcript than towards the 3' end which is expected due to the nature of RNA degradation. QUELLE? Irregularities would appear as individual lines that have slopes differing from the group pattern. The RNA degradation plot overall shows a regular group pattern without slopes that deviate significantly.

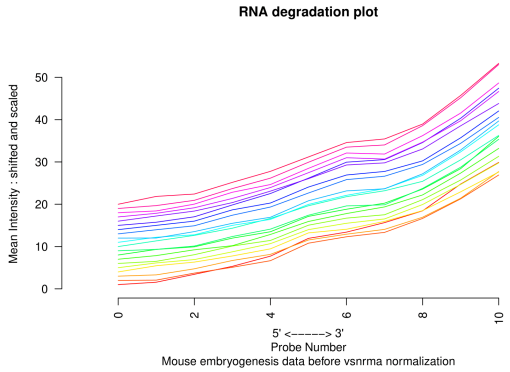

Figure 1: Figure 1.2: RNA degradation plot This plot is used to verify the quality of RNA. Slopes with a higher steepness compared to the overall pattern indicate low quality RNA.

Normalisation

The data had to be normalised before conducting further calculations necessary to obtain differentially expressed genes (limma). Variance stabilising normalisation (VSN) is specific for Affymetrix GeneChip probe level data and thus a reasonable choice for this step. VSN does probe-wise background correction and between-array normalisation. The function returns an ExpressionSet that can be used for further analysation. In the boxplot with VSN-normalised data it stands out that the mean values of the different chips are similar and the intensity differences equalised. Yet, it should be noted that even after normalisation, there is still a slight increase visible in the means of chip E12.5 and E14.5.

Scatter Plot

Furthermore, scatter plots can be used to review the quality of data. We plotted the normalised chips against each other according to the following pattern: E7.5_1 against E7.5_2,.., E8.5_3 against E 9.5_1,... Plots comparing the same measurement points do not show scattering. Some plots comparing different measurement points however display skewed scattering.

MeanSdPlot

To verify the previous normalisation in regards to variance-mean dependence a MeanSDPlot was used on the normalised data. A horizontal line would indicate no dependence. In the plot below the line slightly ascends with increasing mean value. However this is still in range, so no data was excluded.

3 Results

3.1 Analysis of TRA gene expression

In order to determine expression levels of TRAs, the aforementioned data was run through statistical devices and then visualised with different methods. Within the framework of this project, the data was first explored with k-means clustering, applying elbow method, silhouette method, and _____ to determine the optimal number of clusters. This was followed by limma analysis with p values 0.001 and 0.005 on TRA

and Chemokine data and yielded respective tables. TRA and Chemokine expression were plotted to show expression development in different tissues. Some significant plots will be shown here, whereas consecutive data can be found in supplementary material.

3.1.1 Dimensionality Reduction and Clustering (PCA, UMAP, t-SNE)

Dimensionality reduction, including Principal Component Analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP) served as additional quality control.