Final Report

18.07.2022

Clara Certa, Yaxin Chen, Linda Kaupp, Alewtina Towara

Supervisor: Dr. Maria Dinkelacker, Dr. Carl Herrmann

Tutor: Ian Dirk Fichter

Data Analysis for students of Molecular Biotechnology

Heidelberg University

Contents

Table of contents 1

1.Abstract 2

2.Introduction 3

3.Methods and Results 4

3.1 Quality Control and Normalization of the Data 4

setting working directory and loading 4

Single chip control: reading in the microarray chips 4

Boxplots 4

Density Plots 4

RNA Degradation Plot 4

Scatter Plot 4

Data Clean-up 4

3.2 Annotation 4

3.3 Exploratory data analysis 4

3.4 Data analysis 4

3.4.1 Principle Component Analysis 4

3.4.2 Limma analysis 5

3.4.3 Gene set enrichment analysis 6

3.4.4 Venn diagram 6

chemokine 6

Table of contents

1 Abstract

1.Abstract

2.Introduction

3.Methods and Results

###load library

3.1 Quality Control and Normalization of the Data

setting working directory and loading

Single chip control: reading in the microarray chips

First step of the Quality Control is the single-chip analysis by which major quality problems like fingerprints, local irregularities, imprints of pipette tips and extreme light intensities . Therefore we looked at the images of each microarray chip.

In our date set ### Normalization

Boxplots

Density Plots

RNA Degradation Plot

Scatter Plot

Data Clean-up

3.2 Annotation

We want to annotate all the TRA genes in our chips with TRA and ensembl gene information

There are 160 genes which are in the TRA table but not in Ensembl table. We downloaded and checked the Ensembl table again, which did not help.

3.3 Exploratory data analysis

We want to visualize the distribution of the TRAs in our chips according to their max tissue, using a boxplot.

shitty heatamp, do we want to keep it?

3.4 Data analysis

3.4.1 Principle Component Analysis

Using PCA to analyse our 18 chips, to see whether there are any outlier-chip and whether the replicates from different stages cluster with each other.

4 PCs are enough to show more than 90%

We cannot observe any outlier replicate and we can see four cluster (Cluster 1:1-cell,2-cell and 4-cell stage, Cluster 2: 8-cell stage, Cluster 3:Morula, Cluster 4:Blastocyst)

Now we want to use Kmeans on the first 4 PCs (which explain 90% of the Variance) to find potential clusters and compare it with the cluster that we have observed in PCA just now seems like k=2 is the best

Elbow plot suggests 2 Clusters (Cluster 1: 1-cell stage, 2-cell stage and 4-cell stage, Cluster 2: 8-cell stage, morula and blastocyst)

Plot the Siloutte plot

it seems like that k=4 is a good choice

Silouette Plot suggests 4 Clusters: Cluster 1:1-cell,2-cell and 4-cell stage, Cluster 2: 8-cell stage, Cluster 3:Morula, Cluster 4:Blastocyst)

Using Kmeans based on the original expression data (human.vsnrma.df2) to cluster and compare the results with the Kmeans results based on 4PCs

seems like k=2 is the best

Elbow plot suggests 2 Clusters (Cluster 1: 1-cell stage, 2-cell stage and 4-cell stage, Cluster 2: 8-cell stage, morula and blastocyst)

Plot the Silhoutte plot

it seems like that k=2 is a good choice

Silouette plot suggests 2 Clusters (Cluster 1: 1-cell stage, 2-cell stage and 4-cell stage, Cluster 2: 8-cell stage, morula and blastocyst)

Finding genes which explain for the variance by looking at genes with the highst loading in PCs

Use PCA to plot 3 replicates of 1-cell stage and 8-cell stage

two PCs are enough to show 90% variance

Finding variance genes which explain the variance of PC1 between 1-cell stage and 8-cell stage

three PCs are enough to show 90% variance

Finding variance genes which explain the variance of PC1 between 8-cell stage and morula stage

Use PCA to plot 3 replicates of 8-cell stage and blastocyst stage

Use PCA to plot 3 replicates of morula stage and blastocyst stage

four PCs are enough to show 90% variance

Finding variance genes which explain the variance of PC1 between morula stage and blastocyst stage

3.4.2 Limma analysis

Using Limma analysis, we want to find out differential expressed genes (DEGs) between different stages. We will conduct the test among every two stages but only look at the results of the stages between different clusters (showed in PCA): stage 1 cell-8 cell, stage 8 cell - morula, stage morula - blastocyst

creating fit tables between every two stages

using toptable function to create limma table (only significant DEGs!) for every two stages

we annotate the limma table (only significant DEGs) with ensembl

we extract the TRAs from ensembl annotated limma table (only significant DEGs) and annotate these significant differentially expressed TRAs with TRA information

Tissue plot we want to visualize the expression of TRAs among the different stages

###tissue plot facetting

filter DE genes & volcanoplots we create volcanoplots to visualize the DEGs

so far, our limma table only contains significant DEGs. To demonstrate other genes that do not show differential expression we need to create a limma table with all genes. By doing that we found that the ensembl table has some transcripts that repeats itself, so we only select the unique transcripts from the ensembl data

generate a function with the output of limma tables with sig. genes and non sig. genes

anotation of the limma table (complete) with emsembl information

we want to plot three volcanoplots (between 1-cell and 8-cell stage, 8-cell stage and morula stage, morula stage and blastocyst stage). We decided to highlight 10 highst up-regulated TRAs and 10 highst down-regulated TRAs.

Repeating the same procedure with the other three stages

3.4.3 Gene set enrichment analysis

3.4.4 Venn diagram

Using Venn diagram, we want to find out which genes are continuously up/down-regulated during every stage. They could be important for the embryogenesis.

chemokine