MICROARRAY DATA ANALYSIS

DMYTRO PRAVDYVETS[1]

12.06.2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Knowing that intercellular bacteria Chlamydia has some sort of effect on gene expression of different gene in human epithelial cells we are going to test the gene expression in control and infection cells to see if there is really different expression or not, which may help further analysis of the topic. After the analysis genes are going to be annotated to see their function.

* *BDBI, ESCI-UPF*
[1] *Omics techniques, Alex Sanchez Pla*

1 OBJECTIVES

The main objective of this analysis is to find differential expressed genes. The data for this experiment was obtained based on reprogramming of host cells during chlamydial infection. To identify biological processes polysomal mRNA as a proxy of actively transcribed mRNA.

2 METHODS

In this section datastructure and methods that were used for the study are going to be mentioned

1. Data structure Data of this experiment is structured in this way: Infection of total RNA, control total RNA, Infection polysomal RNA, control polysomal RNA
2. Basic plots andPCA for data validation For this sort of analysis we do boxplot, clustering and PCA plots to check how the samples are distributed
3. Quality analysis Using arrayQualityMetrics package from Bioconductor we assess reproducibility, identify apparent outlier arrays and compute measures of signal-to-noise ratio of the data
4. Data normalization and filtering
   - Using Robust Multi-Array Average Expression Measure we normalize the observations to get rid of artifacts
   - Filter the data by discarding low signal observation from the expression set using nsFilter, which discards 40213 genes
5. Repeat boxplot, clustering and PCA analysis for the normalized data The new graphs are fine
6. Creating design and fitting linear model
   - Design and Fitting data into a linear model. For this we use 1 variable matrix where we join control, infection, total RNA and only polyA RNA into 4 different combinations and create a matrix where rows are the samples

Sample	InfectionTotal	ControlTotal	InfectionPolyA	InfectionTotal
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	1	0
10	0	0	0	1
11	0	0	0	1
12	0	0	0	1

   - And creating a contrast matrix based on this comparisons

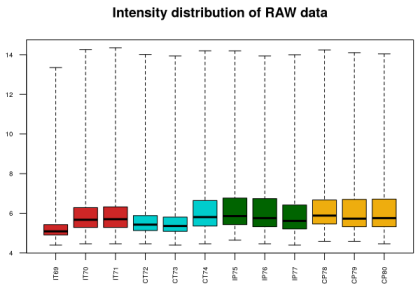
Levels	itvsct	itvsip	itvscp	ctvsip	ctvscp	ipvscp
InfectionTotal	1	1	1	0	0	0
ControlTotal	-1	0	0	1	1	0
InfectionPolyA	0	-1	0	-1	0	1
ControlPolyA	0	0	-1	0	-1	-1

7. Obtaining the most deferentially expressed genes
   - Creating top table for each comparison of data type. Basing on comparison of expression of same genes in different situations we create a table for each comparison to see the most differential expressed genes
   - Joining 5% of mostly expressed genes of each top table and creating 1 with the mostly differential expressed genes. Out of all the top tables we take 5% of the most differential expressed ones and join them into one top table that will be used for plots and annotations
8. Volcano plot of the genes to visualize the difference Volcano plot of this genes Done for better visualization of the data
9. Annotation of those genes Annotating the selected genes using Bioconductor

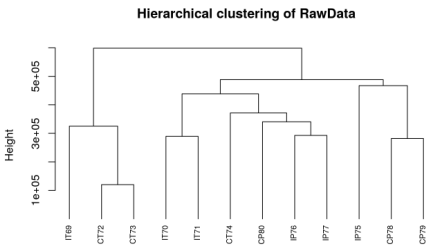
3 RESULTS AND DISCUSSION

3.1 Raw data

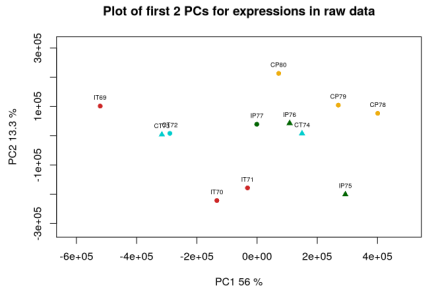
After performing the analysis on the raw data we can inherit from the boxplot that the data is more or less equally destributed without any type of artifacts, which is quite good, also clusters that are seen on the cluster plot seem to go on with the boxplot (samples with lower values in boxplots are together). PCA analysis supports the previously said information and also we can see that the difference between groups in raw data is pretty big



(a) Boxplot raw data



(b) Cluster raw data

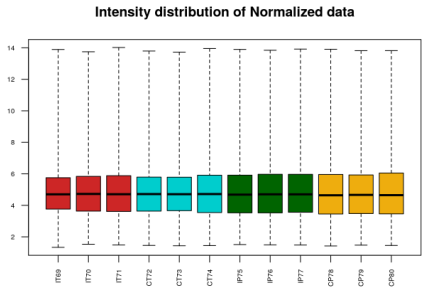


(c) PCA plot raw data

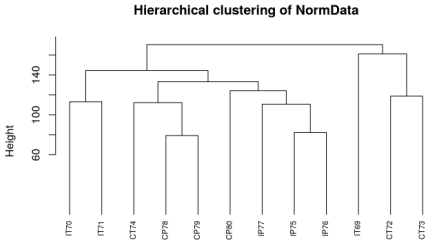
Figure 1: Raw data plots

3.2 Normalized data

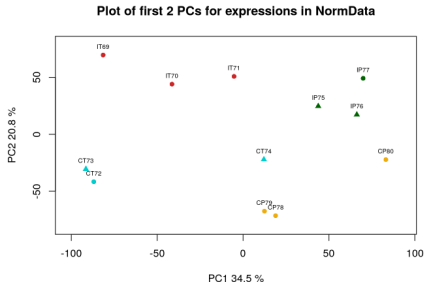
Normalization of data is an option in this case just to get rid of few extreme values that might have been errors. One of the things that appeared somewhat strange is how the boxplots now are vey very simimlar, but this is pretty normal after normalization, clustering is now also different for example node 74 and 80 are not together now, 75 and 76 are not longer in the same cluster neither, but besides that other nodes are stil the same. Also PCA analysis plot is now has less differences between the samples than before due to normalization, of course. Basically what we inherit from this is that this normalized data can be used for further analysis



(a) Boxplot normalized data



(b) Cluster normalized data

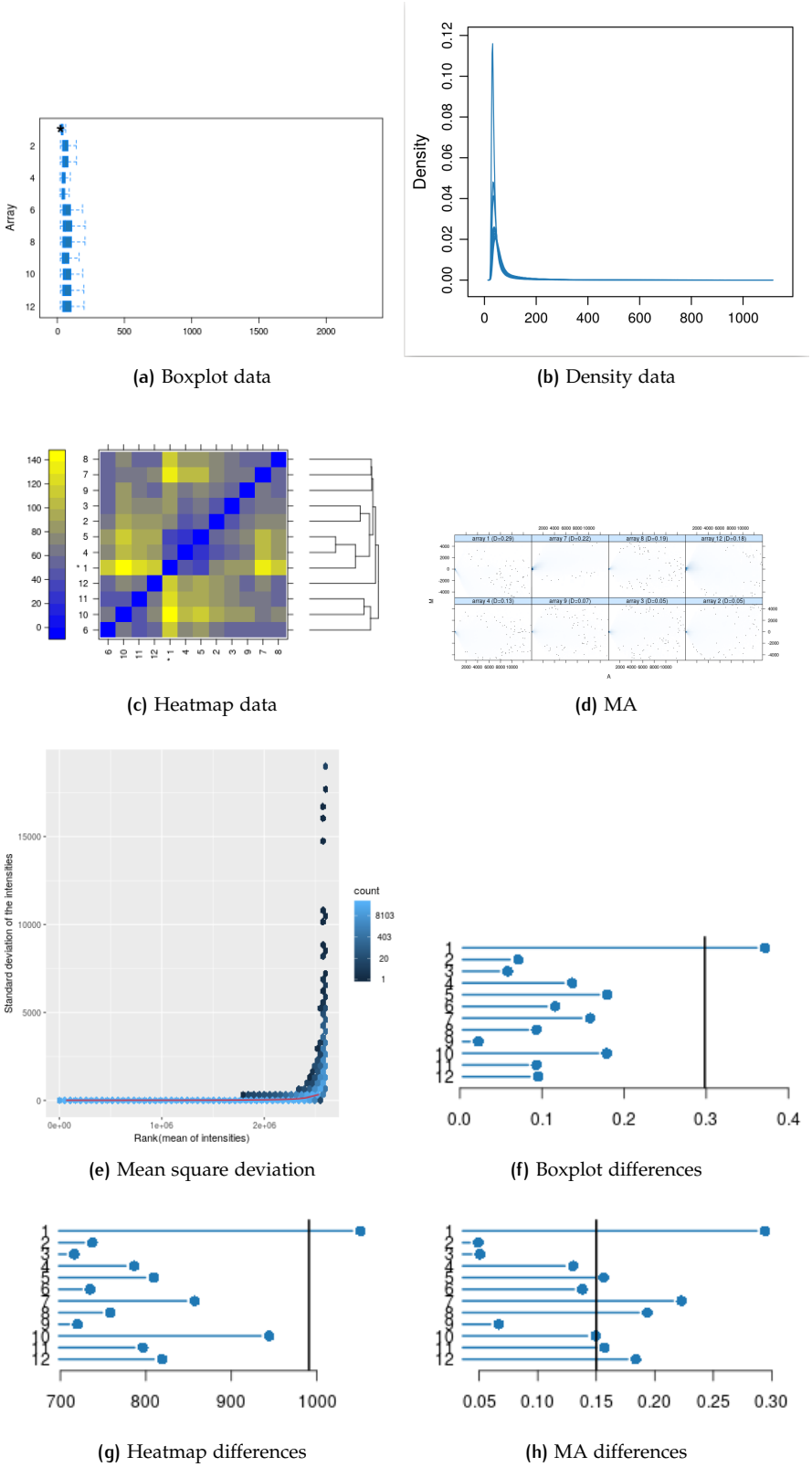


(c) PCA plot normalized data

Figure 2: Raw data plots

3.3 Array quality analysis and filtering

Using **nsFilter**, which discards 40213 genes we obtain a smaller expression set which is easier to use for further analysis. One of which is Array quality analysis. Out of this analysis we can confrm the information said in the raw data analysis and tha the data is usable without anything strange in it as signal-to-noise ratio of the data is fine and no huge outliers are found.



(a) Boxplot data

(b) Density data

(c) Heatmap data

(d) MA

(e) Mean square deviation

(f) Boxplot differences

(g) Heatmap differences

(h) MA differences

Figure 3: Array quality analysis plots

3.4 Find differential expressed genes

Using the matrices of design described in the Methods we fit the observations into a linear model and account for Bayes. The resulting model is going to be used for

top tables of the genes. After creating a top table for each comparison we join the 5% of all the tables into one.

For a better visualization of data we do a volcano plot of the joined top table

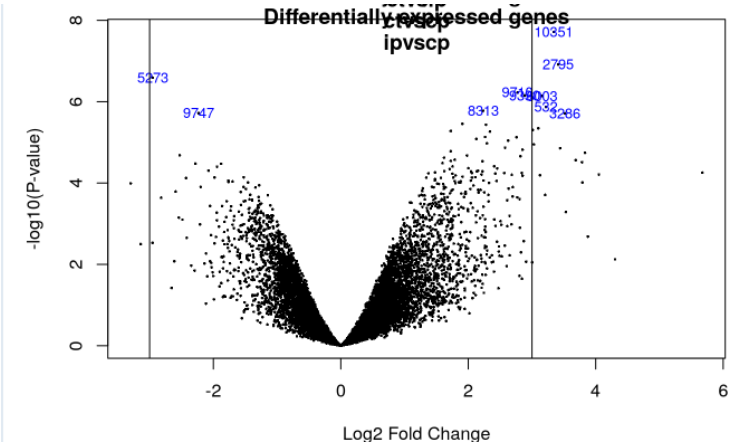


Figure 4: Volcano plot

This plot shows us that there are some genes that have a quite different expression even in the set of mostly expressed genes which may indicate that those genes expression is changed when polyA RNA or total RNA are infected, to find this out we trace the value of the gene and search for it in the top tables. Also to see which are the samples that manifest differential expressed genes we do a heatmap of the top table in which we can see that infected samples manifest more differential expressed genes compared to control

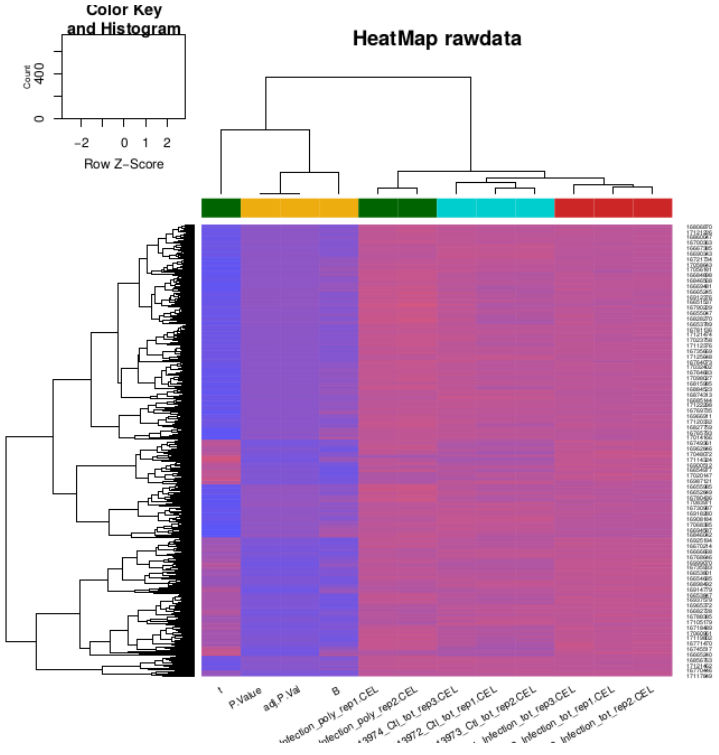


Figure 5: Volcano plot

3.5 Gene annotation and function

After annotating the genes from the top table we can see what genes are they and what function they are involved in. This list is pretty big here are just few examples:

- 2-oxoglutarate and iron dependent oxygenase domain containing 1
- 3-hydroxyisobutyrate dehydrogenase
- 5-hydroxymethylcytosine binding
- 5'-nucleotidase ecto
- ubiquitin specific peptidase 22
- vacuolar protein sorting 4 homolog B
- zinc finger BED-type containing 8
- zyg-11 family member A
- uroporphyrinogen III synthase
- speedy/RINGO cell cycle regulator family member E17
- solute carrier family 20 member 1

There are many more functions, that are stored in a file **functions.txt**

3.5.1 *Final decision*

After finishing the analysis we can state that there is a clear pattern of different expression in quite a few genes that are analysed and that the most of those genes come from the infection in polyA RNA. Knowing this information further studies of the infection specificated in polyA RNA can be done.