MAT 5900 Final Project: Analysing time-course data

Yun-Chyau, Chiou

Introduction

In this project, I will analyze the time-course RNAseq data with DESeq2(Michael I. Love and Huber 2018) and Rpackage called "moanin"("Moanin: An r Package for Time Course RNASeq Data Analysis" 2015), both to filter the Differential expression gene in the time-course dataset. DESeq2 is a package using likelihood ratio test that compares how well a gene's count data fit a "full model" (with independent variables, like time) compared to a "reduced model" (without those variables). On the Other hand, "moanin" is developed to provide a simple and efficient workflow for long time-course gene expression data. We could apply this "moanin" to do differential expression between conditions based on either individual condition comparisons per time point, or by fitting spline functions per gene. After doing both differential expression gene analysis, I will then extract intersection union(differential expression gene) to do the clustering.

In the following project, I will follow the following main steps. (Prinz 2020)

- Input data
- Quality control & Normalization
- Find differentially expressed genes
- Clustering
- Pathway enrichment

Overview of the Dataset

This is data from a micro-array time-course experiment, exposing mice to three different strains of influenza, and collecting lung tissue during 14 time-points after infection (0, 3, 6, 9, 12, 18, 24, 30, 36, 48, 60 hours, then 3, 5, and 7 days later). The three strains of influenza used in the study are (1) a low pathogenicity seasonal H1N1 influenza virus (A/Kawasaki/UTK4/2009 [H1N1]), a mildly pathogenic virus from the 2009 pandemic season (A/California/04/2009 [H1N1]), and a highly pathogenic H5N1 avian influenza virus (A/Vietnam/1203/2004 [H5N1]. Mice were injected with 105 PFU of each virus. An additional 42 mice were injected with a lower dose of the Vietnam avian influenza virus (103 PFU). (al. 2015)

data: is a data frame with 39544 rows corresponding to genes and 209 corresponding to samples. The rownames give the RefSeq name of the gene. [figure 1]

meta: is a data frame with 209 rows corresponding to samples and 3 named columns (the first column is just an index from 0-208) [figure 1]

- **Group** The treatment group of the sample. "C"=Control, "K"=Kawasaki strain, "M"=California strain, "VH"=Vietnam strain, "VL"=Vietnam at lower dosage (103 PFU).
- **Replicate** Identifies the replicate – each combination of treatment and timepoint was replicated three times (except for VH at timepoint 3, which has only 2 replicates).

- **Timepoint** Identifies the time passed (in days) since infection of the sample

	GSM1557140	GSM1557141	GSM1557142
NM_009912	4.364173	4.338212	4.467095
NM_008725	8.227778	8.318347	8.763661
NM_007473	4.656177	5.010798	5.134367
ENSMUST00000094955	5.845917	5.631973	5.689894
NM_001042489	10.527553	10.391772	10.545010
NM_008159	5.178836	4.961242	4.968316
NM_001013813	6.070302	5.980675	5.702378
AK039774	6.048065	6.005902	5.931749
NM_013782	8.817178	8.969646	8.919900
NM_028622	4.904111	4.851039	4.821741
NM_028007	9.018037	9.001353	9.010408
NM_198093	8.175626	9.323055	9.663536
NM_145123	5.020457	4.906300	5.282160

		Group	Replicate	Timepoint
GSM1557140	0	K	1	0
GSM1557141	1	K	2	0
GSM1557142	2	K	3	0
GSM1557143	3	K	1	12
GSM1557144	4	K	2	12
GSM1557145	5	K	3	12
GSM1557146	6	K	1	18
GSM1557147	7	K	2	18
GSM1557148	8	K	3	18
GSM1557149	9	K	1	24
GSM1557150	10	K	2	24
GSM1557151	11	K	3	24

Figure 1: Data dataframe and Meta Dataframe

Quality Control

This data set has already done normalization, as a result, I will focus on the quality part. In the time-course data quality control, there are two ways, firstly, we need to filter the lowly expressed data. Secondly, we can create a correlation plot between each samples. For the first steps, there are several ways to filter out lowly expressed genes. I have tried the CPM way which means we filter out the genes if counts-per-million (CPM) above 0.5 in at least two samples; however, the result turns out that it did not filter out any genes since every CPMs are all above 0.5. Then I tried to filter out based on the median absolute deviation, according to the distribution plot (fugure 2 left hand side), it is reasonable to filter out the median absolute deviation >0.5. (Ziebell 2021) (Belinda Phipson 2020)

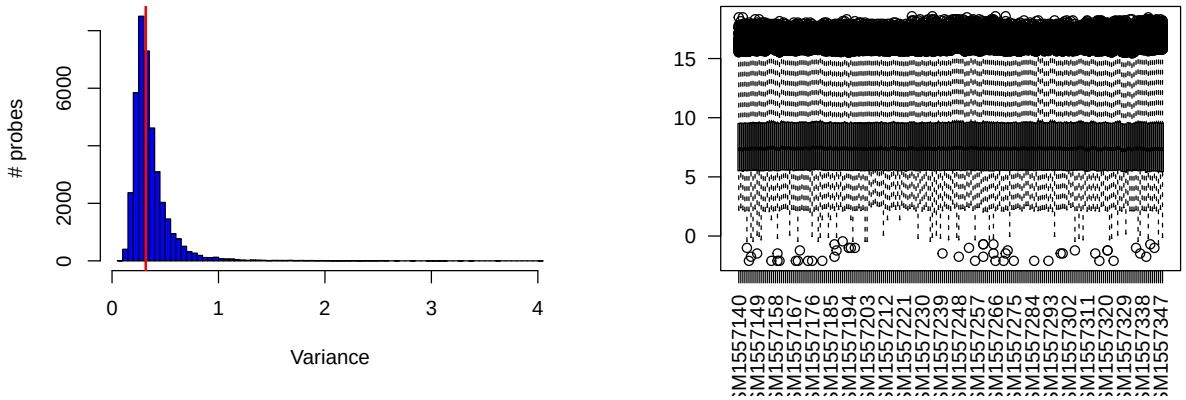

Figure 2: median absolute deviation distribution plot

The right hand side if Figure 2 plot is showing the box plot of the the expressed gene, and according to the plot, we could check that after filtering out the gene, the distribution seems more normal and stable, which means there is not much different between each genes.

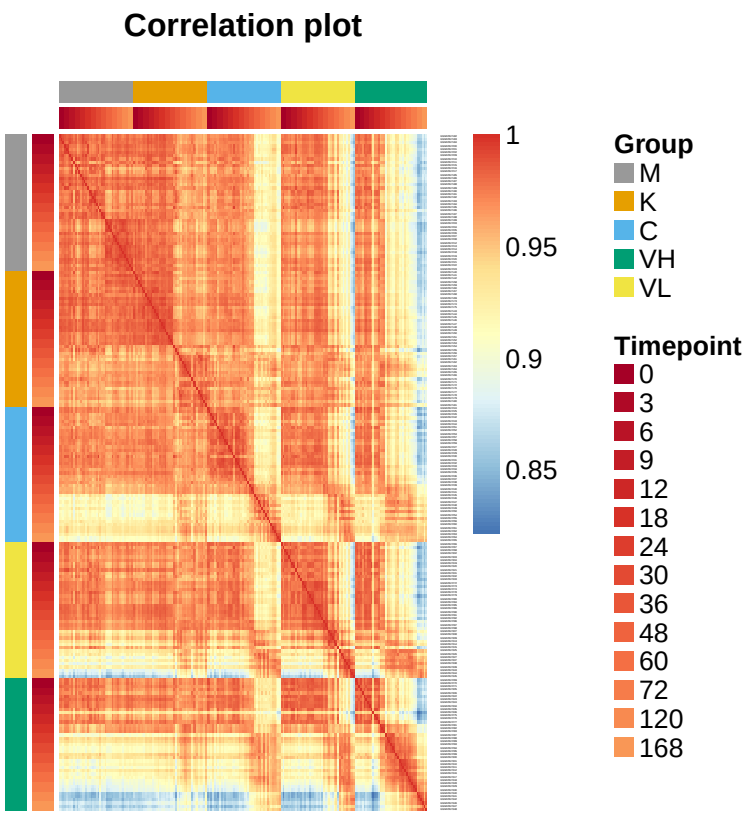

Figure 3: heatmap correlation plot

According to the correlation plot, it shows some interesting patterns. First of all, the 120 and 168 time-point of the Vietnam strain samples are the one that are the most different from control samples. Secondly, the less pathogenic the strain is, the closer the samples are to the control condition.

Differential expression analysis of time-course data

Time-course analyses with DESeq2

DESeq2 is a package using likelihood ratio test (LRT) to analyze the time-course data. The LRT test can be especially helpful when performing time course analyses. We can use the LRT to explore whether there are any significant differences in treatment effect between any of the time point. The following is the statistic of the LRT. For this test two models are estimated per gene; the fit of one model is compared to the fit of the other model. where $m1$ represents the reduce model, and $m2$ represents the full model. (SARA 2021) (Jihe-Liu 2020)

$$LR = -2ln(\frac{m1}{m2})$$

Our full model ($m2$) is:

$$m2 = \beta_3 replicate + \beta_2 treatment + \beta_1 t + \beta_0,\ t = timepoint$$

Our full model ($m1$) is:

$$m1 = \beta_3 replicate + \beta_2 treatment + \beta_0,$$

We can know if the gene's expression fits a pattern of decrease or increase over the different time points. In analyzing the data, DESeq2 take estimates dispersion of each gene's expression by taking into account the dispersion of genes expressed at similar levels. There are biological replicates for each time point, they will also be calculated dispersion.

After applying the DESeq2 to the time-course data, there are 9935 out of 39544 genes are filtered out for differential expression analysis. According to the artical: Time-Course Analysis of Gene Expression During the Saccharomyces cerevisiae Hypoxic Response (Nasrine Bendjilali 2017), I will use another package "moanin" to also filtered out for another group of differential expression analysis, and then extract intersection union to do the clustering.

Time-course analyses with moanin: Weekly differential expression analysis

DE analysis in moanin.(Nelle Varoquaux 2016) Moanin is a package in R that could provide functionality for performing both pre-time point analysis and global analysis. Here we focus on global analysis where we consider the expression pattern globally over time, and consider what genes have either different patterns between conditions or a changing pattern (i.e. non-constant) over time. We fit a spline model to each gene, and use that model to test different kinds of different expression across time. We create a moanin-model below, and the following is the summary of the moanin model.

```
## Moanin object on 209 samples containing the following information:
## Group variable given by 'Group' with the following levels:
##  M  K  C VL VH
## 42 42 42 42 41
## Time variable given by 'Timepoint'
## Basis matrix with 35 basis_matrix functions
## Basis matrix was constructed with the following spline_formula
## ~Group + Group:splines::ns(Timepoint, df = 6) + 0
##
## Information about the data (a SummarizedExperiment object):
## class: SummarizedExperiment
## dim: 39544 209
## metadata(0):
## assays(1): ''
## rownames(39544): NM_009912 NM_008725 ... NM_010201.1 AK078781
## rowData names(0):
## colnames(209): GSM1557140 GSM1557141 ... GSM1557347 GSM1557348
## colData names(5): '' Group Replicate Timepoint WeeklyGroup
```

One type of DE analysis we can do is to compare different treatments to each other, for every time point. This method is from limma for the DE analysis. In the following workflow, I create contrasts to compare control groups verses other groups, which means in the following groups, I do the contracts between control groups vs Kawasaki strain, control groups verses California strain, control groups verses Vietnam at lower dosage. The results give the raw p-value (_pval), the FDR adjusted p-values (_qval).

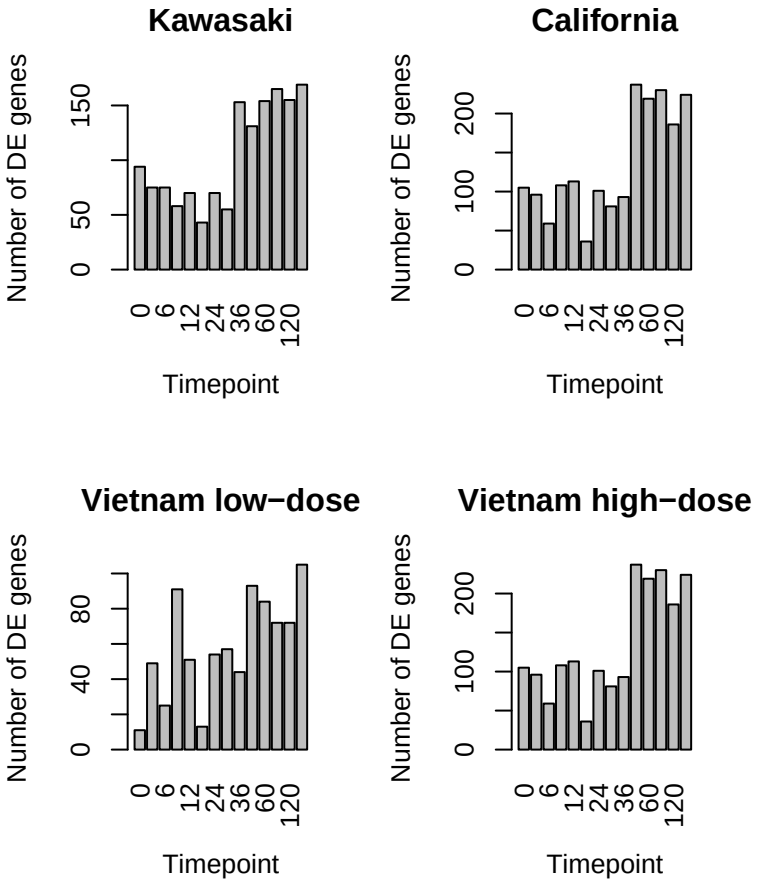

Figure 4: Number of DE genes while comparing control group to other treatments

Figure 4 shows the amount of the DE genes while comparing the control group to other treatments based on the moanin model we create in the previous workflow. According to the plot, we could see that there are some weird patterens in the plot, for example, the 18H of California strain, it suddenly dropped from 12H to 18H. While there may be biological differences at those time points for some genes, it seems unlikely that the large majority of genes differentially expressed at timepoint 12H stop being differentially expressed at 18H and then jump back to being differentially expressed at 24H. A more likely explanation is that there are some technical or biological artifacts about the samples for 18H that are creating higher variation and thus less ability to detect significance.

It can be challenging to compare the data across weeks because the outcome of the weekly differential expression analysis may be quite low and various weeks may have a varied number of replicates. As an alternative, we can fit a smooth spline to each gene in each group and then use differential testing to see whether there are any differences between each groups. Furthermore, in the moanin package, we could do log fold change with the function "estimate_log_fold_change". In the function, we could choose different ways for the log_fold_change, including timely, timecourse, sum, max and min. Here, I apply the method of sum, which gives the sum of the absolute difference in the means across time points.

After doing all the aanalysis process via moanin, there are 3811 genes to keep, which means 3811 genes are differentialy expressed.

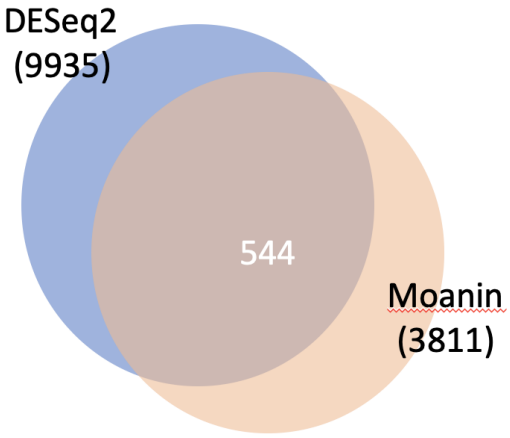

Figure 5: Interect Union DESeq2 and moanin

Clustering

The adaption of the k-means clustering:

- Spline estimation: For each gene, fit a spline function to the basis of your choice.
- Rescaling of splines: Rescales the estimated spline function for each gene so that the values are between 0 and 1, allowing comparisons between genes.
- K-Means: Applies k-means to the rescaled fitted values of the splines to estimate the centroids of the clusters.

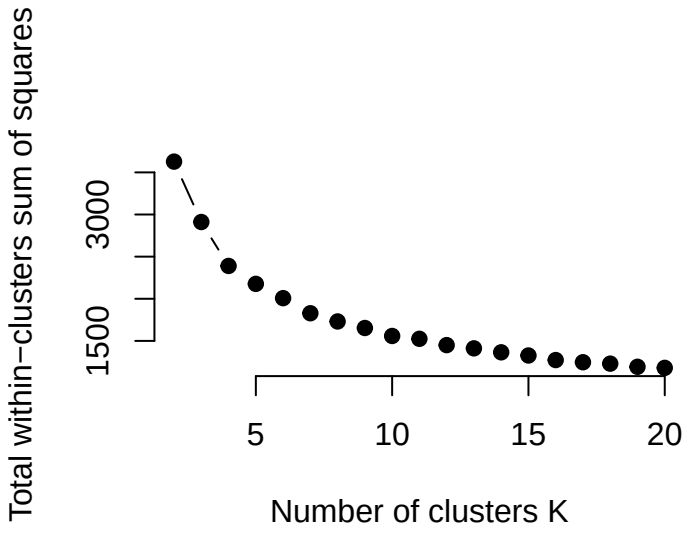

Figure 6: k-means clustering of DE genes

The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. According to figure 6, I will choose k=5 for cluster number, because from k=5, the curve becomes flat and stable. (DeGraaf 2020)

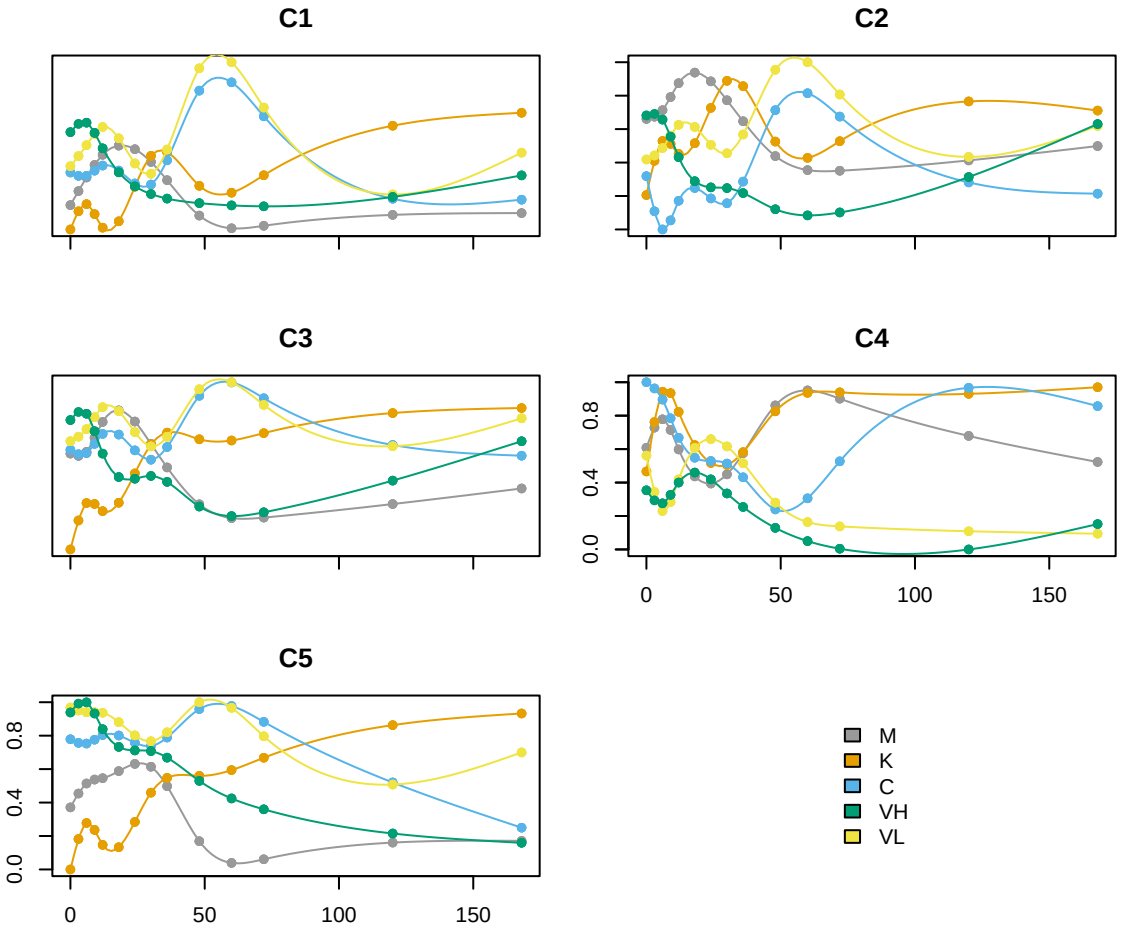

Figure 7: k-means clustering of significant expression gene

Clusters	Number_of_genes
1	127
2	122
3	103
4	71
5	91

According to figure 7, we could see that there are some similar patterns in different clusters. The intersting fact is that VH and VL have similar pattern, which means higher dosage or lower dosage won't affect the result. Based on this clustering, we could predict in different group while in these genes of the lung tissue during the timepoints. So that could predict how will it become in the future. Now I would like to look into the cluster 2, since it contains genes with strong differences between the different influenza treatments and the control.

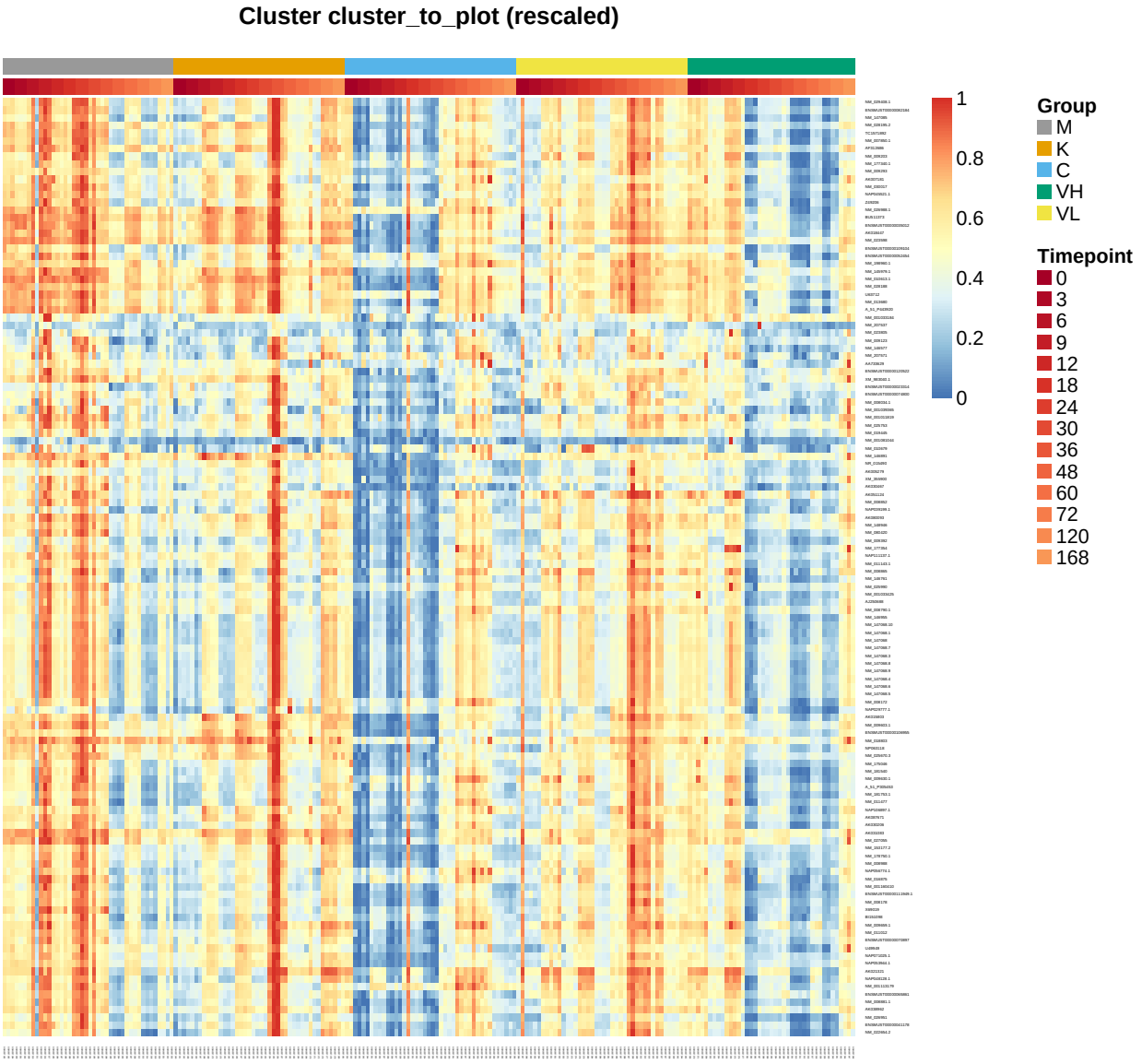

Figure 8: Heatmap for cluster 2

According to figure 8, gene NM_207537, NM_001081044 has the special comparing to other genes in every timepoint and treatments. And in cluster 2, we conclude that California Strain has higher lung tissue during time-point 24H and 30H. Vietnam at lower dosage and CalCalifornia Strain has similar pattern in this cluster.

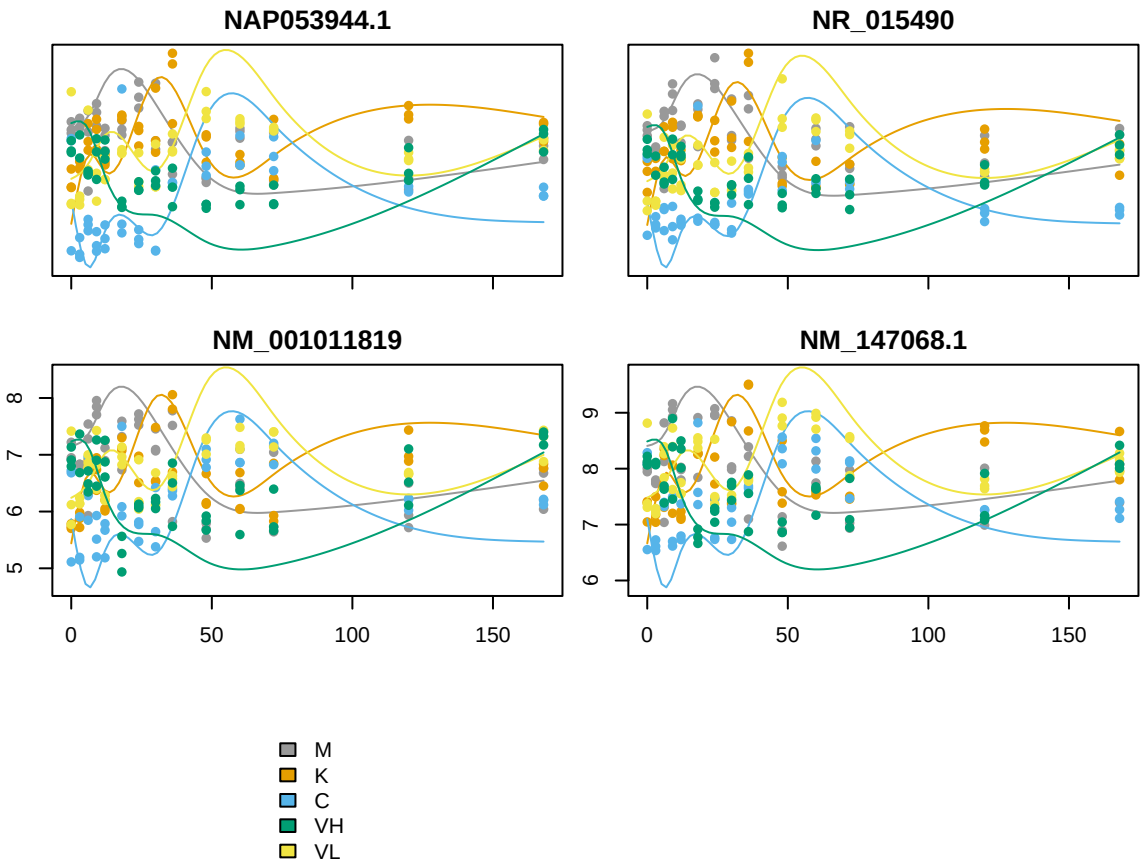

Figure 9: Top 4 significant genes for cluster 2

GO enrichment analysis

After we have the clusters, we can now do the GO enrichment analysis, which perform enrichment analysis on gene sets. First of all, we need a gene list that contains all the genes we are interested in, here I use the gene in cluster 2, since we already looked into the cluster. And then I use GO enrichment analysis to map the most affecting gene. Here, I applied the topGO package which provides tools for testing GO terms while accounting for the topology of the GO graph. ## GO:0070060, GO:0003947, GO:0000721, GO:0051990, GO:0047305, GO:1902270, GO:1901235, GO:1900749, GO:1901737, GO:1901735, GO:0070626, GO:0047298 GO:0047777. which are related to actin filament nucleation, galactosylglucosylceramide N-acetylgalactosaminyltransferase activity, butanediol dehydrogenase activity, 2-hydroxyglutarate dehydrogenase activity, 3-amino-2-methylpropionate-pyruvate transaminase activity, carnitine transmembrane transport, carnitine transmembrane transporter activity, carnitine transport, mevalonic acid biosynthetic process and mevalonic acid metabolic process.

Conclusion

In conclusion, time-course analysis has shown that different treatments causes widespread and complex changes in gene expression in different timepoint. Based on the cluster patterns, we can predict that for the same gene, how will the gene present during the timepoints while in different treatment. The step for

this project could apply AutoCor method, which is mentioned in (Nasrine Bendjilali 2017) to analyze the time-course data, and filter the intersect union genes for AutoCor and DESeq. So that, the result can be more percise and efficient.

References

al., Shoemaker et. 2015. "Time Course Transcriptomic from Mouse Lung Tissues Infected with Influenza." https://nellev.github.io/timecoursedata/reference/shoemaker2015.html.

Belinda Phipson, Matt Ritchie, Anna Trigos. 2020. "RNA-Seq Analysis in r." https://combine-australia.github.io/RNAseq-R/06-rnaseq-day1.html#Hierarchical_clustering_with_heatmaps.

DeGraaf, Stephanie. 2020. "Time-Course Analysis and Clustering of Gene Expression Data." https://escholarship.org/uc/item/16k3r09c#author.

Jihe-Liu. 2020. "DGE Analysis Using LRT in DESeq2." https://hbctraining.github.io/DGE_workshop_salmon/lessons/08_DGE_LRT.html.

Michael I. Love, Vladislav Kim, Simon Anders, and Wolfgang Huber. 2018. "RNA-Seq Workflow: Gene-Level Exploratory Analysis and Differential Expression." https://bioc.ism.ac.jp/packages/3.7/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html#time-course-experiments.

"Moanin: An r Package for Time Course RNASeq Data Analysis." 2015. https://nellev.github.io/moanin/articles/documentation.html.

Nasrine Bendjilali, Gurmannat Kalra, Samuel MacLeon. 2017. "Time-Course Analysis of Gene Expression During the Saccharomyces Cerevisiae Hypoxic Response." https://academic.oup.com/g3journal/article/7/1/221/6031506.

Nelle Varoquaux, Elizabeth Purdom. 2016. "A Pipeline to Analyse Time-Course Gene Expression Data." https://f1000research.com/articles/9-1447#ref-16.

Prinz, Jeany. 2020. "Analyzing Gene Expression Data." https://www.dataversity.net/analyzing-gene-expression-data/#.

SARA. 2021. "Note for DEseq2 Time Course Analysis." https://r-craft.org/r-news/note-for-deseq2-time-course-analysis/.

Ziebell, Frederik. 2021. "RNA-Seq Quality Control." https://cran.r-project.org/web/packages/RNAseqQC/vignettes/introduction.html.