Differential Expression Analysis in R

Project conducted under Data Fellowship with Numeric Mind

Submitted by:
Sameer Dangol
Srijana Shrestha
2-19-2022

Acknowledgements

We would like to express our sincere gratitude to Numeric Mind for providing us with an opportunity to be a part of the Data Fellowship program.

We would also like to earnestly acknowledge the sincere efforts and valuable time given by the mentors Mr. Binod Jung Bogati and Mr. Sanjay Hamal.

Table of Contents

Abstract

The differential expression of genes between different cell conditions can be measured using RNA-Seq, where the RNA from those cell conditions are extracted and sequenced. Sequencing technologies have allowed us to extract large volumes of data which needs to be pre-processed and analysed to generate meaningful information. We use the different packages available in R to conduct differential expression analysis in the count data of basal and luminal mouse mammary cells of virgin, lactating and pregnant mice. By doing the analysis, we made a Mean-Difference plot (MD-plot) and Volcano plot showing the differentially expressed genes in between basal mammary cells of pregnant and lactating cells. This workflow, with minor tweaks can be used to make plots for other datasets as well.

Introduction

RNA Sequencing (RNA-seq)

RNA-Seq is a technique which is used to measure the transcription of each gene in a biological sample. This is mostly used to analyse the transcriptome, the set of expressed genes, which indicates which of the genes encoded by our DNA are turned on or off and to what extent.

The measurement of gene expression on a genome-wide scale has become common practice due to the advent of next-generation sequencing (NGS) technologies which have lower costs and produce a higher volume of data. Scientists can now use RNA-Seq to simultaneously measure expression of tens of thousands of genes for multiple samples. Initially the cost-limiting factor used to be generating the RNA-Seq data, but today due to advances in the NGS technologies, the cost has shifted to the storage and analysis of the data.

RNA is first isolated from the sample and is sequenced, which produces sequencing reads. These reads are then aligned to a reference genome, so that the number of reads aligned to a reference genome can be counted. This results in a table of counts, where we can perform the necessary statistical calculations in R. While mapping and counting are important tasks in the RNA-Seq workflow, here in the project, we start from the count data and start the analysis.

Data used in the project:

The data used in the workflow comes from a Nature Cell Biology paper, EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival (Fu et al. 2015). The raw data (sequence reads) and the processed data (counts) was downloaded from the Gene Expression Omnibus (GEO) website by searching the accession number GSE60450. The study compared the gene expression in luminal and basal cells harvested from the mammary glands of virgin, 18.5 day pregnant and 2 day lactating mice with 2 mice per stage. So, six groups are present, with one for each combination of cell type and mouse status., with each group containing two replicates. The reads have already been aligned to the mouse genome and the reads mapped to mouse genes have been counted to obtain the counts file. This file was downloaded in the local computer and is the starting point for our analysis.

Methodology

The following operations were carried out for the differential expression analysis. The codes are in the Rmd file sent separately in the email.

1. Installing Packages
2. Data Import and Formatting
3. Converting counts to DGEList object
4. Adding Annotations
5. Filtering Lowly Expressed Genes
6. Quality control
7. Normalisation
8. Differential Expression

Installing Packages

The packages edgeR, limma, Glimma, org.Mm.eg.db, gplots and RColorBrewer were installed and the corresponding libraries were loaded.

Data Import and Formatting

The counts data and the sample information file was imported from the local computer using the read.delim function. Then the data formatting was done by removing the first two columns from the counts data (containing EntrezID and gene length). The EntrezGeneID was stored as rownames variable.

Converting counts to DGEList object

The counts file was converted to a DGEList object using the function DGEList() in the package edgeR. DGEList is an object that stores different parameters in different slots that makes it easier to conduct statistical analyses.

Adding Annotations

The only annotation in the table as of now is the Gene ID, which is not very informative. So, we add some annotation information using the org.Mm.eg.db package, where using the select function, we create a variable ann and add columns, gene symbols and full gene name. Then we add the annotation variable ann to the DGEList object created before.

Filtering Lowly Expressed Genes

Genes with very low counts across all libraries provide little information and interfere with some of the statistical approximations used later. These lowly expressed genes therefore need to be removed. So, here we calculate the counts-per million (CPM) values and retain the genes if they are expressed at a CPM above 0.5 in at least two samples. The cpm() function in the edgeR library was used and then filtering was done.

Quality control

After storing the counts in a DGEList object and filtering lowly expressed genes, we made a few plots to check the quality of the data. We checked the library size, i.e., the number of reads for each sample. We also plotted barplot of library sizes, boxplots of logCPMs, Multidimensional Scaling Plots and heatmap to visualise the relationship between samples.

Normalisation

There is composition bias in the data due to differences in the sizes of the libraries. So, normalisation needs to be performed. We use the calcNormFactors() function to apply TMM normalisation which gives us the relative RNA production levels from RNA-seq data.

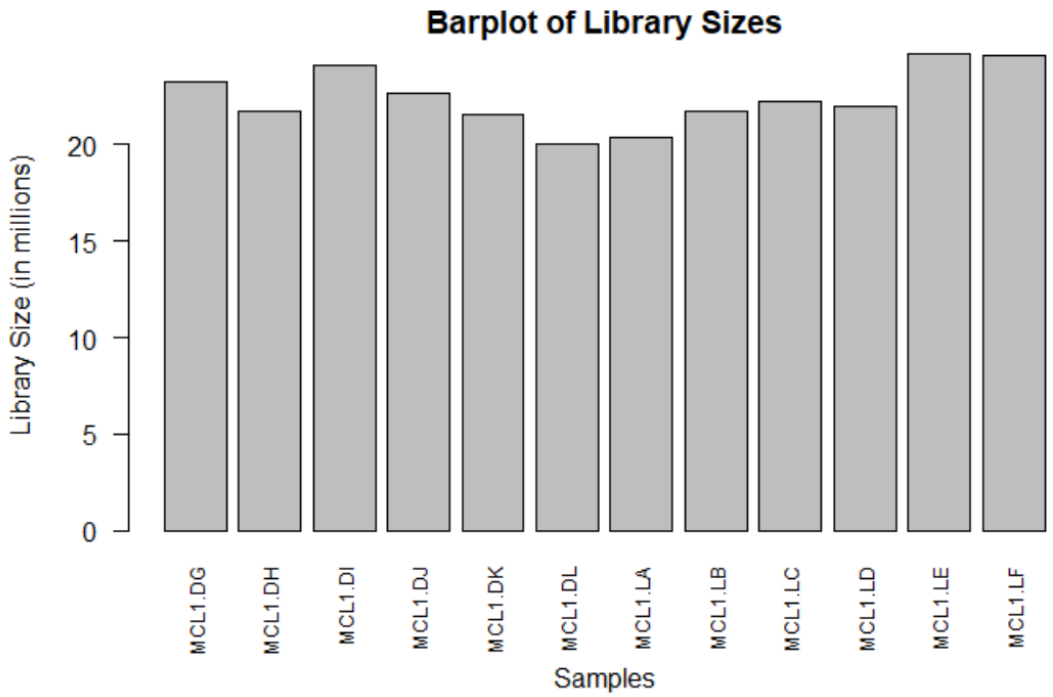
Differential Expression

After checking the quality and normalising the data, we can start testing for differentially expressed genes. We create the design matrix to test differences in status in the different cell types. Here, we want to know which genes are differentially expressed between pregnant and lactating in basal cells only. Then using the voom() function in the limma package, we transform the read counts into

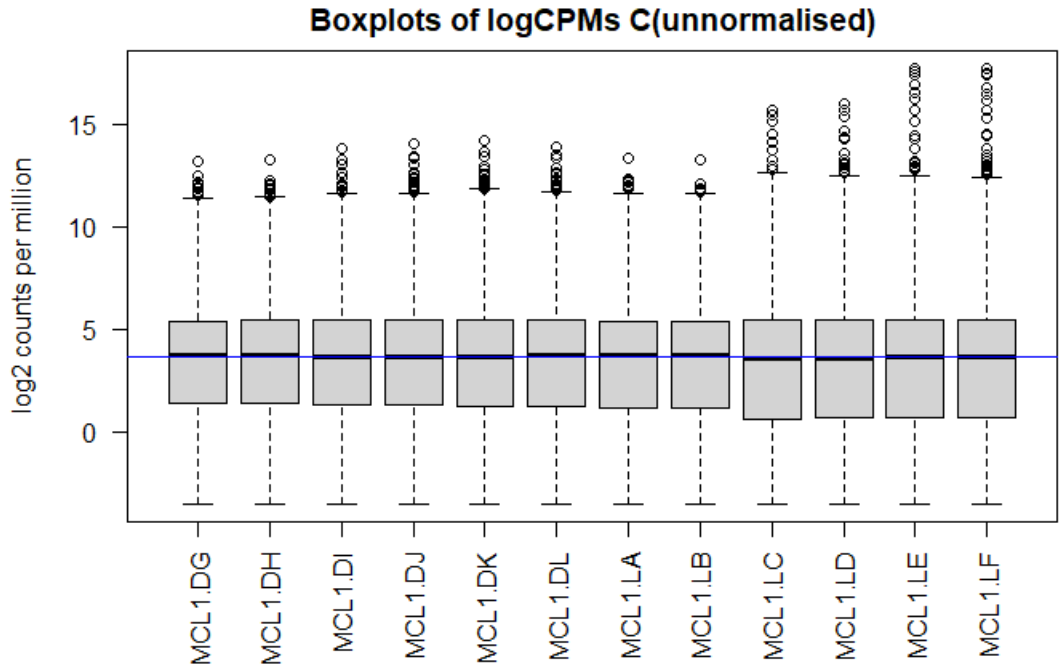
logCPMs while taking into account the mean-variance relationship. Then, standard limma commands(lmFit, makeContrasts, contrasts.fit, eBayes, etc) were used on the voom transformed data to test for the differentially expressed genes. The lmFit function needs the voom object and the design matrix and fits a linear model for each gene. makeConstrasts function was then used to specify the comparison we need to make. The linear model and the contrasts were used by the contrasts.fit function which gave the statistics and the estimated parameters of the comparison. These statistics are then visualised by making MD plots and volcano plots.

Plots made in Rstudio:

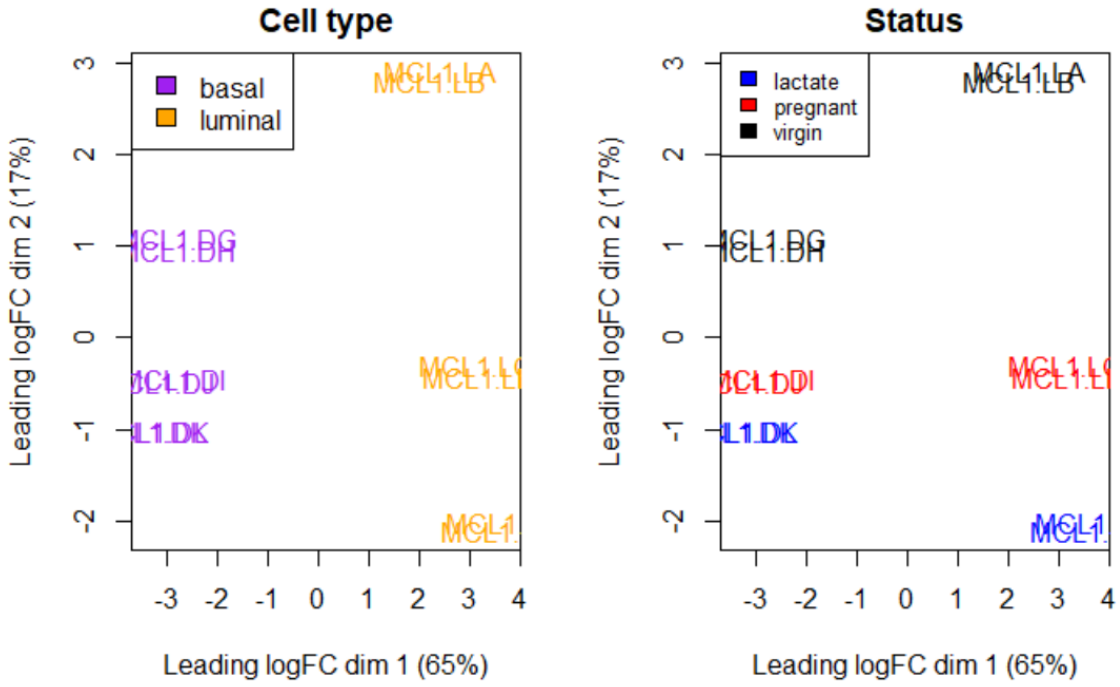
Barplot



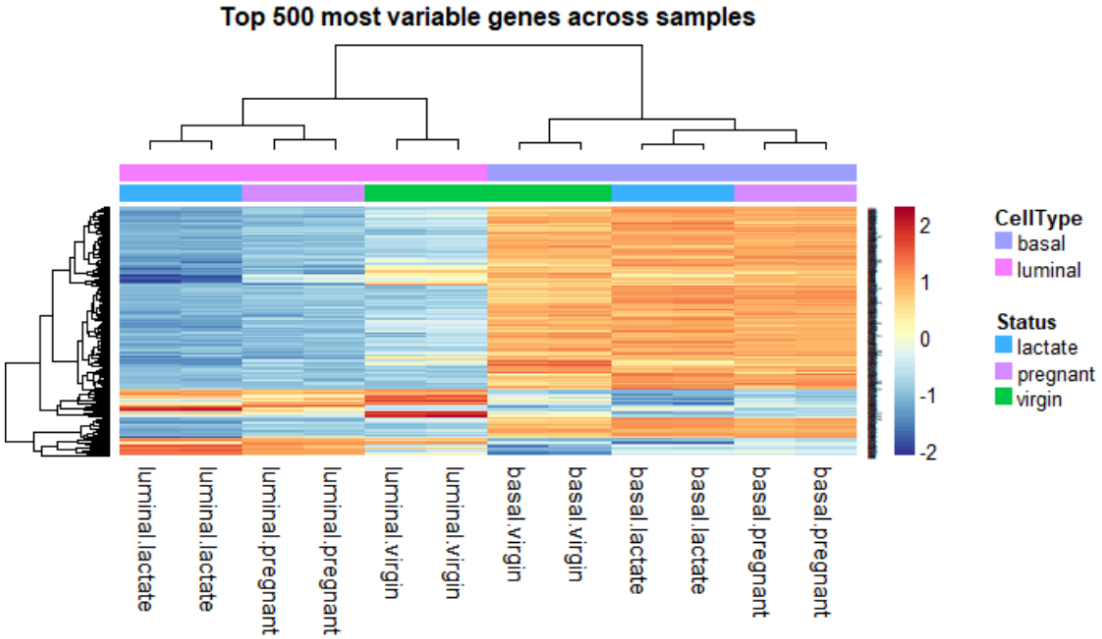
Boxplot



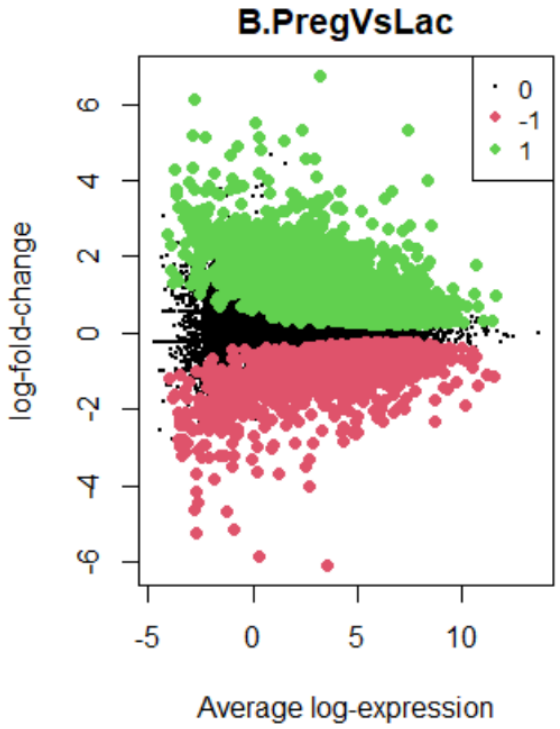
MDS Plot



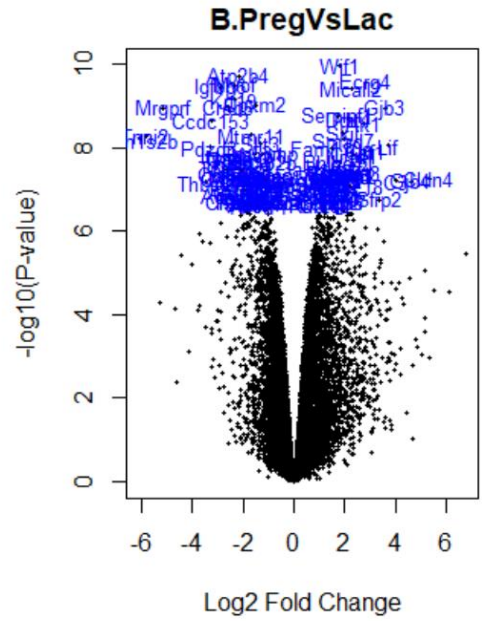
Heatmap



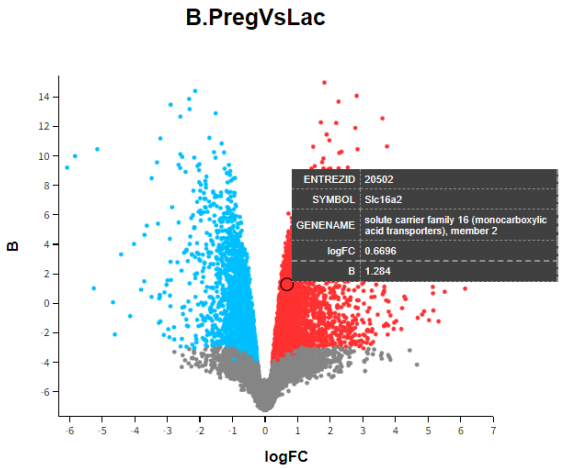
MD Plot

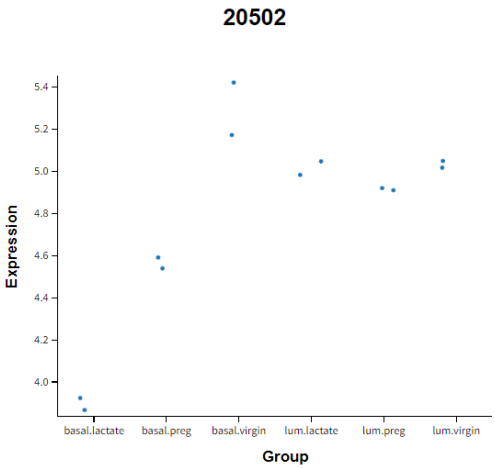


Volcano Plot



Interactive Volcano Plot





Search:

ENTREZID	SYMBOL	GENENAME	logFC	B
11302	Aatk	apoptosis-associated tyrosine kinase	0.3349	-4.798
11303	Abca1	ATP-binding cassette, sub-family A (ABC1), member 1	0.2468	-5.595
11304	Abca4	ATP-binding cassette, sub-family A (ABC1), member 4	1.217	-3.626
11305	Abca2	ATP-binding cassette, sub-family A (ABC1), member 2	0.3722	-5.195
11306	Abcb7	ATP-binding cassette, sub-family B (MDR/TAP), member 7	-0.09397	-6.69
11307	Abcg1	ATP binding cassette subfamily G member 1	0.5679	-5.121
11308	Abi1	abl interactor 1	-0.2643	-3.156

Conclusion

We used R to analyse and visualise the genes differentially expressed between pregnant and lactating basal cells from mice using published scientific data. This pipeline gave us a foundational understanding on how to conduct differential expression on similar datasets.

References

1. Fu, N. Y., Rios, A. C., Pal, B., Soetanto, R., Lun, A. T., Liu, K., Beck, T., Best, S. A., Vaillant, F., Bouillet, P., Strasser, A., Preiss, T., Smyth, G. K., Lindeman, G. J., & Visvader, J. E. (2015). EGF-mediated induction of MCL-1 at the switch to lactation is essential for alveolar cell survival. Nature Cell Biology, 17(4), 365–375. https://doi.org/10.1038/ncb3117

2. Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2018). RNA-seq analysis is easy as 1-2-3 with Limma, GLIMMA and edger. F1000Research, 5, 1408. https://doi.org/10.12688/f1000research.9005.3

3. RNA-seq analysis in R - github pages. (n.d.). Retrieved February 18, 2022, from https://combine-australia.github.io/RNAseq-R/06-rnaseq-day1.html

4. RNAseq data analysis in R - Notebook. RNAseq data analysis in R - notebook. (n.d.). Retrieved February 18, 2022, from http://monashbioinformaticsplatform.github.io/RNAseq-DE-analysis-with-R/RNAseq_DE_analysis_with_R.html