Bioinformatics: Week 7 & 8 Deliverables

Ashlesha Patil

Construction and Analysis of a ceRNA network Reveals Potential Prognostic Markers in Colorectal Cancer

- The paper aims to **screen and identify potential prognostic biomarkers in colorectal cancer**
- The ceRNA network is built upon the idea of a **potential cross-talk between the miRNA and the lncRNA**, which both compete for their binding sites on mRNA and therefore regulate and control post-transcriptional gene expression and can contribute to disease development
- Nine hub genes were screened and further in-depth analysis showed that the **MFAP5-miR-200b-3p-AC005154.6 axis** as a potential prognostic marker

Pipeline of the paper, compared to our analysis

Similarities to what we have done:

- Screened their genomic data and identified the differentially expressed genes by **setting a criteria for logFC and the adjusted p-value**
- Identified potential cellular functions of the DEGs, using the **KEGG pathway analysis**
- Most of cellular functions of screened genes indicated roles in **tumorigenesis**

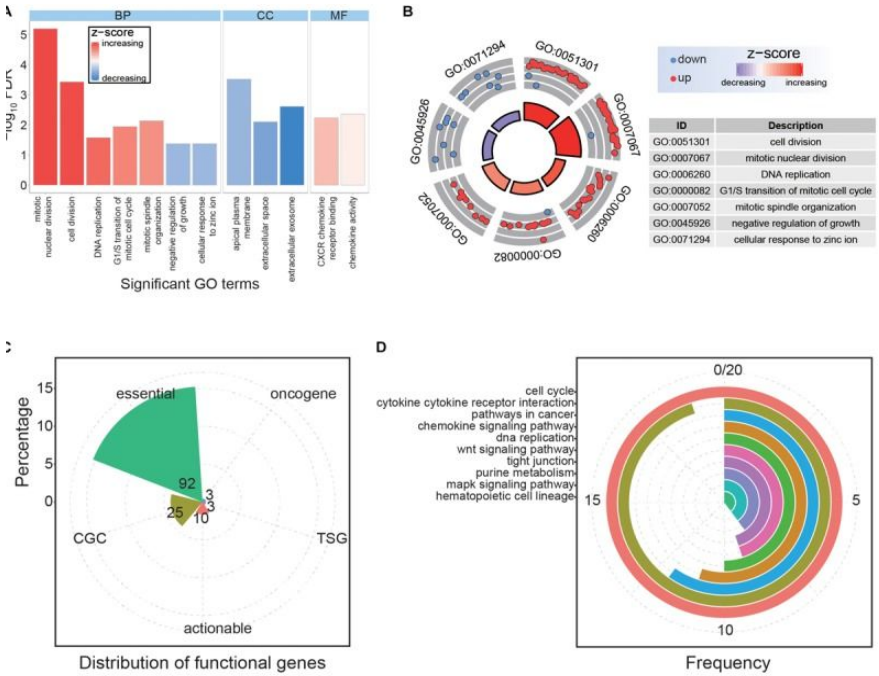


Figure 2A,B,C and D, showing functional analysis of DEGs (Li Guo et al., 2020)

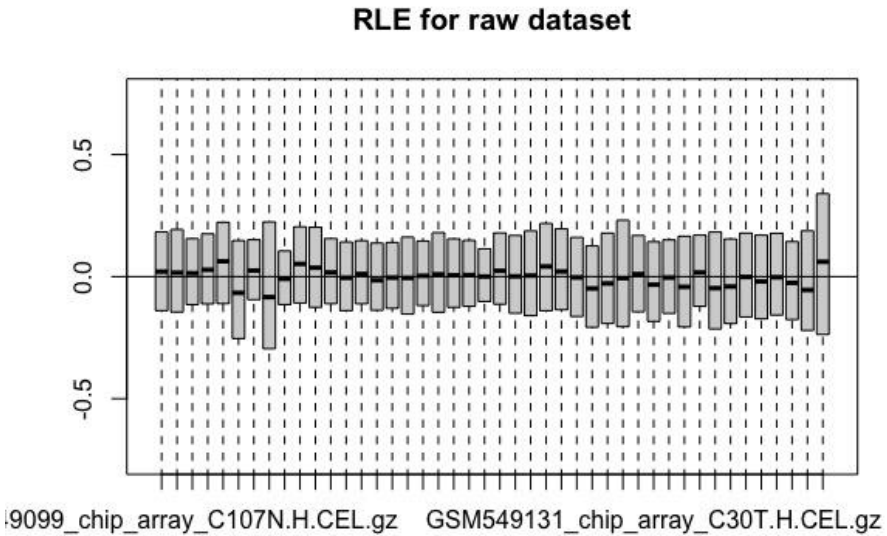
Pipeline of paper, compared to our analysis

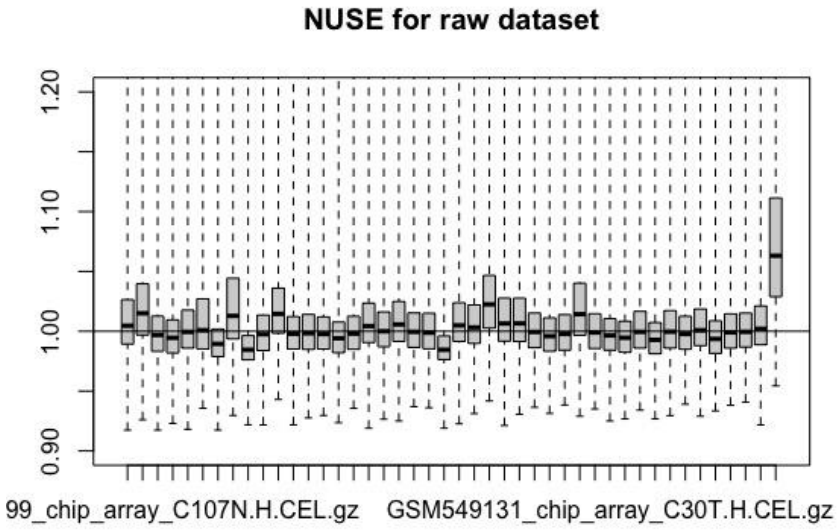
Differences, and steps not mentioned in the paper:

- Did not mention about how DEGs were obtained before selection criteria, including **Quality Control, Normalisation and gene filtering,** which was performed in our analysis
- The paper went further than functional analysis, screening hub genes based on **PPI network** (protein-protein interaction) and also looking at **related interacting miRNA and lncRNA to construct the ceRNA network**

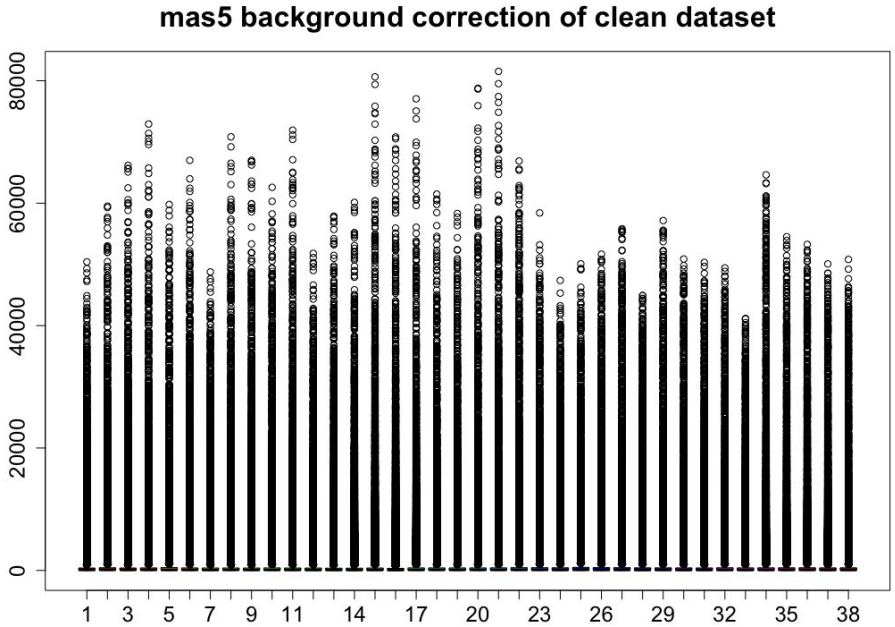
GSE21510: Homogenised control and cancer microarray datasets of colorectal cancer

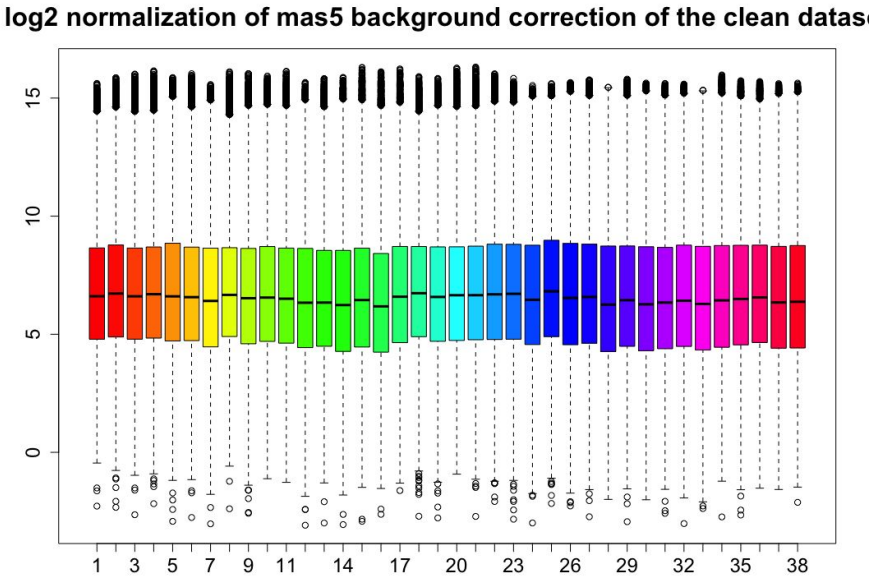
RLE and NUSE (Raw Dataset)





Normalisation: mas5 and log2



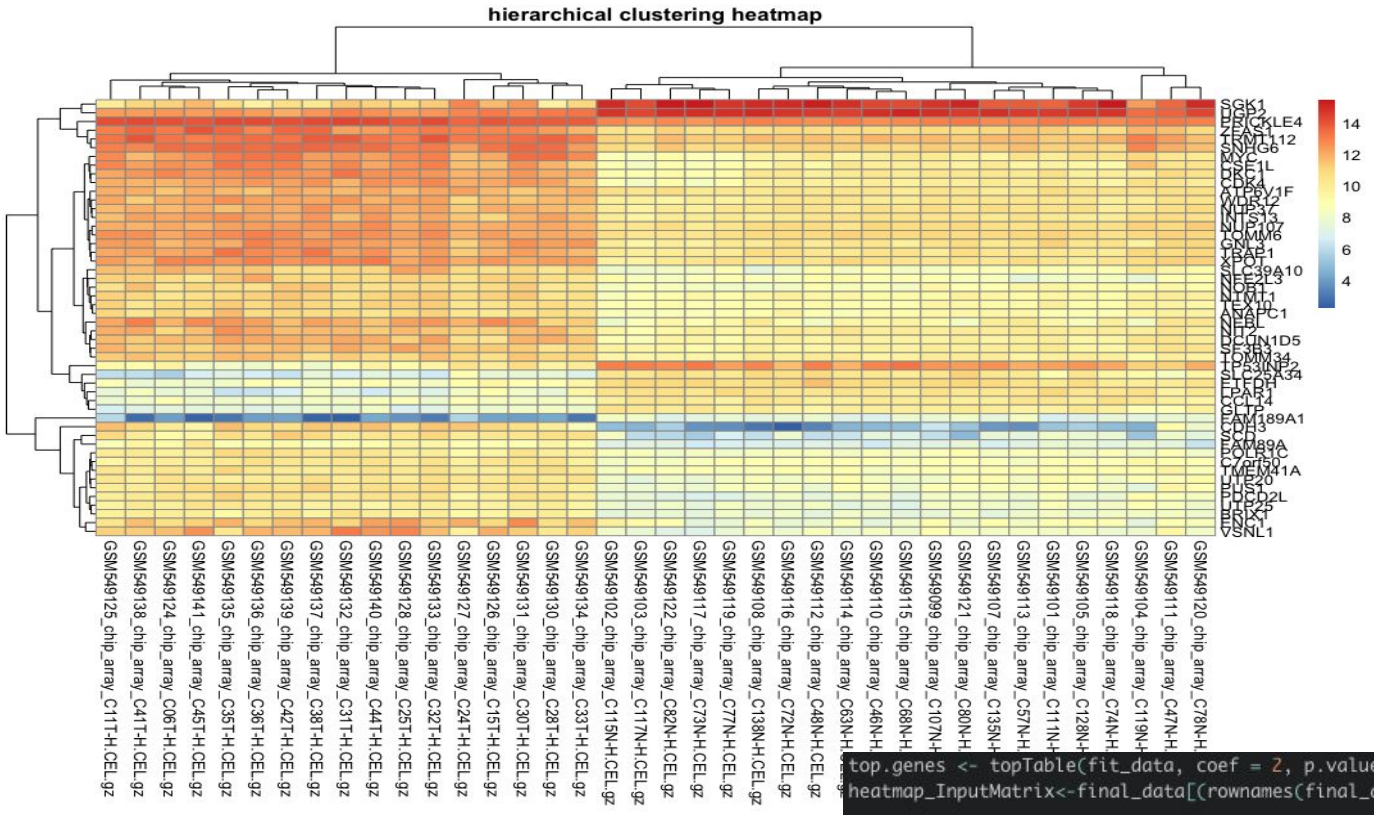


Annotation, Gene Filtering and Limma Analysis

	logFC	AveExpr	t	P.Value	adj.P.Val	B
NFE2L3	2.396832	9.693982	15.65288	2.422957e-18	2.240955e-14	31.60797
CCL14	-1.758813	9.274195	-15.47791	3.532127e-18	2.240955e-14	31.24263
FAM189A1	-3.899680	5.865376	-15.37328	4.431661e-18	2.240955e-14	31.02257
BRIX1	1.905679	8.757824	15.37092	4.454515e-18	2.240955e-14	31.01758
DKC1	2.539452	11.105302	15.02496	9.508884e-18	2.851288e-14	30.28118
VSNL1	3.480855	9.696249	14.99890	1.007279e-17	2.851288e-14	30.22518
UGP2	-1.997386	13.789786	-14.98394	1.041200e-17	2.851288e-14	30.19298
ATP6V1F	1.425505	11.075306	14.89458	1.269553e-17	2.851288e-14	30.00017
CDK4	2.565324	10.820776	14.89257	1.275237e-17	2.851288e-14	29.99583
FAM89A	1.997316	8.055343	14.72550	1.851817e-17	3.464542e-14	29.63287

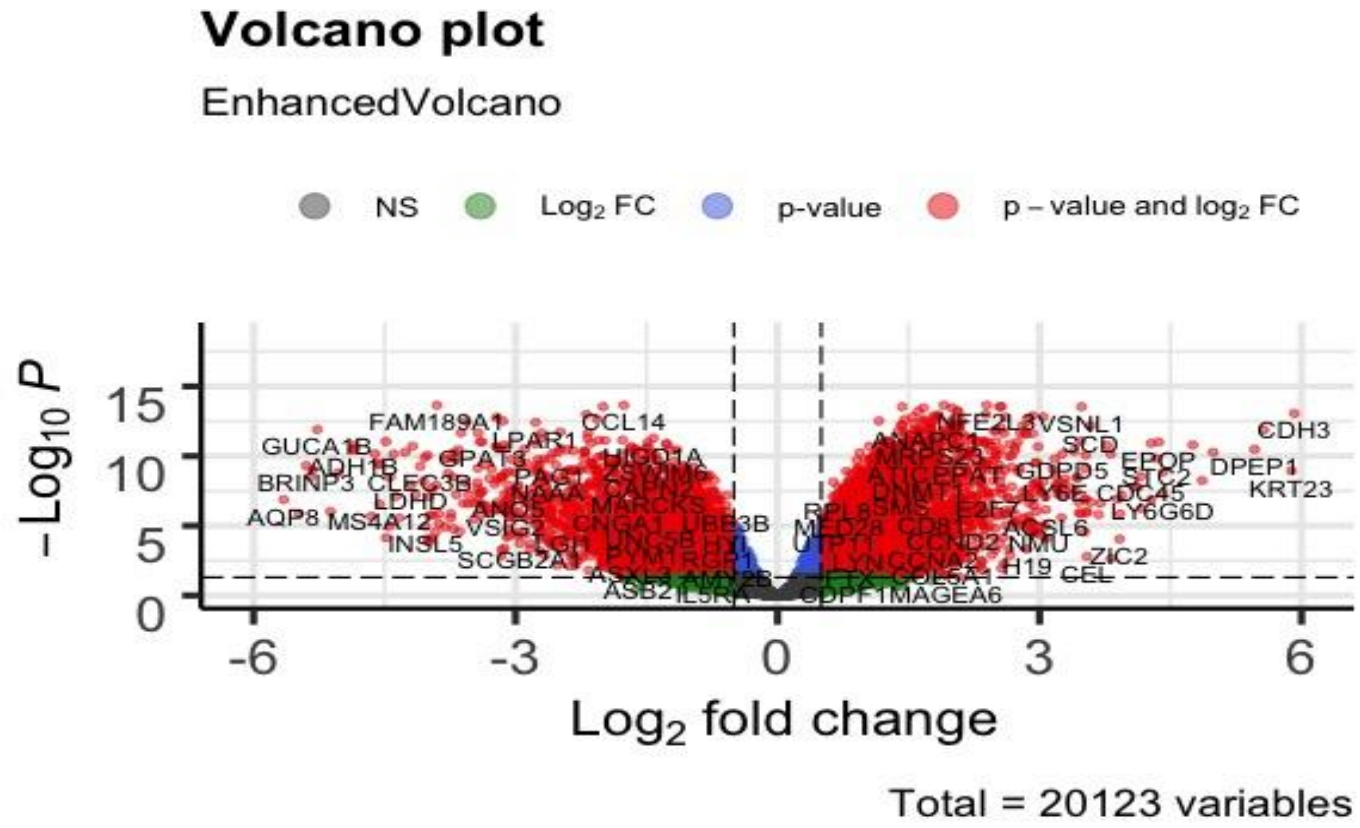
```
#gene-Filtering

mean_all <- apply(final_data, 1, mean)

final_data <- final_data[mean_all > quantile(mean_all,0.04, na.rm = T),]

# analyse with Limma
library(limma)

sampletype <- factor(c(rep("control",21),rep("cancer",17)))
group <- model.matrix(~factor(sampletype,levels = c('control','cancer')))
fit_data <- lmFit(final_data,group)
fit_data <- eBayes(fit_data)
output <- topTable(fit_data, adjust.method = "fdr", sort.by = "B", number = 20123)

final_output <- output[1:100, ]

write.csv(final_output,file = "Final output with top 100 data.csv")
```

Heatmap of Differentially Expressed Genes

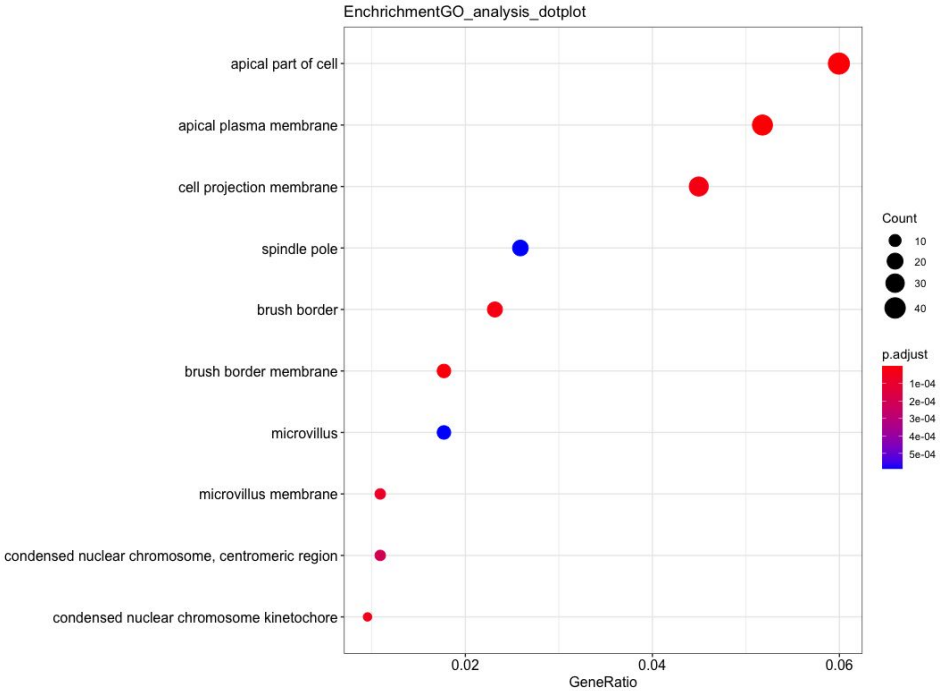


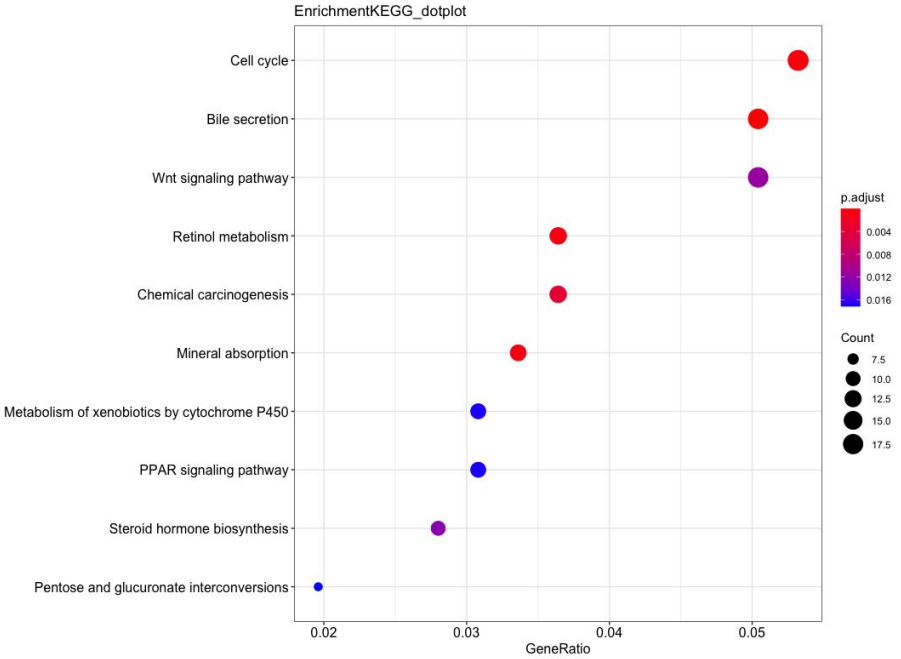

```
top.genes <- topTable(fit_data, coef = 2, p.value = 0.05, number=50)
heatmap_InputMatrix<-final_data[(rownames(final_data) %in%  rownames(top.genes)),]

library(pheatmap)
pheatmap(heatmap_InputMatrix, main = "hierarchical clustering heatmap")
```

Volcano Plot of Differentially Expressed Genes



Functional Analysis





Things learned over the internship

- How to analyse large scale genomic data, using R and also the GEO database
- Interpreting my analysis and applying it back to the paper and understanding how it fits in with the main purpose of the paper
- Working and interacting with my team and sub-teams on a virtual platform
- Connecting with my teammates and networking on LinkedIn, building a better professional network.

Challenges Faced (final deliverables and project)

- While working on the final deliverables, I firstly used the whole dataset of GSE21510, which was quite large and created a few errors. One of the leads told me to use just the homogenised tissue samples, and that helped a lot in my further analysis.
- Overall in the project, there were some issues my sub-team (Team Magic) faced, mostly in the technical areas, but we always set up meetings every week and worked through the deliverables together.

Three Achievement Highlights

- Working with the Bioinformagic Team and my sub-team (Team Magic) and communicating, interacting and learning together on a virtual platform
- Building on my networking and expanding my professional network
- Applying my statistical knowledge in R from my previous classes to a biological context

Bioinformatics and next steps!

- This internship helped me gain a much better understanding in the field of bioinformatics, in which I did not have any prior understanding before
- Learning the whole pipeline and the analysis that is going into understanding these large scale genomic datasets and the results and interpretations that you can get from them has really fascinated me and I am excited to carry on and apply everything that I have learnt onto my further research in my area of study in neuroscience