**Normalisation Strategies for Differential Gene Expression Analysis**

Method

Dataset

In order to compare all normalisation strategies, we ran each of them on the same dataset. From GEO…… this paper….

We eliminated LPS data from the sample because we are Interested in the developmental stages of microglia. Since the simple size is 47 and 10 batches are present.

Computational steps

1. Distribution of samples pre and post normalization
2. Loading and filtration of lowly expressed genes

edgR function was used

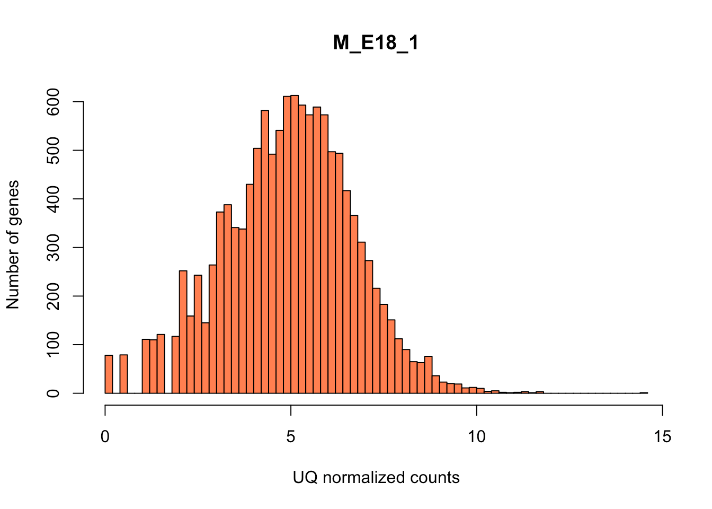
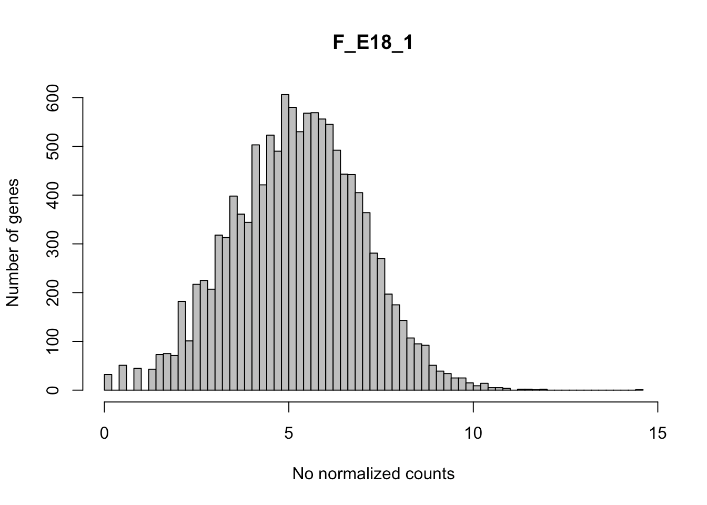
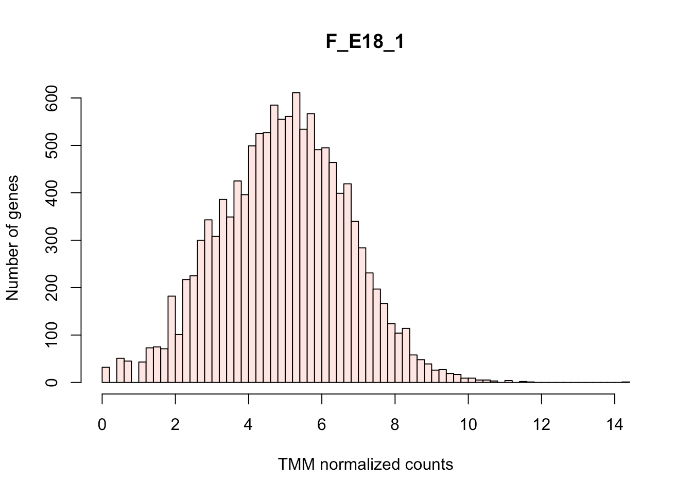
1. Checking for the Batch effect on data
2. Normalization of sequencing depth by CPM and log transformation
3. Normalization by TMM, TMMwsp, UQ and RLE and a negative control
4. Getting the Differentially Expressed Genes (DEGs) from all the methods
5. Venndiagram of the DEGs

**Results**

All the four methods have same statistical assumption, majority of genes are not differentially regulated, this holds true for the data set we are using. (more detail and cite),

1. **No significant differences in the distribution of the samples**

No significant differences were observed in the samples using different normalization techniques as an example the plot for F\_E18\_1 (Female embryonic day 18 replicate 1) sample is shown below as an example. The distribution is near to Normal distribution, hence, for downstream differential analysis parametric tests would be suitable. The peak of curve is at around 5 where is mean, median and mode lies. (Code with all the histograms of all the samples are in Supplementary file 1: [Distribution\_plots.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Distribution_plots.html?csf=1&web=1&e=6GEQTY))

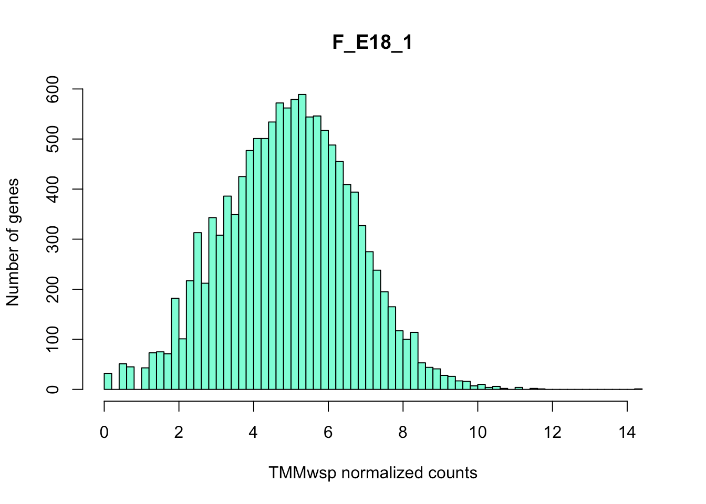


A

B

B

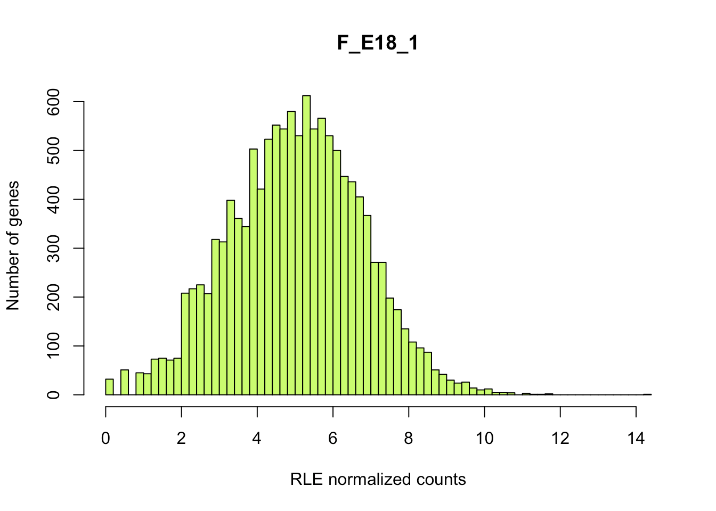
A



C

D

Fig 1: Histogram of a sample F\_E18\_1 after various normalization strategies. It depicts relation of number of genes on y-axis and normalised count values on x-axis. A) No any normalization method is used raw counts are CPM and log transformed B) TMM normalization method C) TMMwsp normalization method D) UQ normalization E) RLE normalization



E

C

1. **Non normalized data had slight fluctuation of median values than other methods**

Except for box plot of counts which is not normalized, in all other four approaches median and upper quartile values are similar. This shows that other four approaches are making the samples more comparable than no normalization. (code and analysis are found in Supplementary file 2 :[normalisation\_strategy.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/normalisation_strategy.html?csf=1&web=1&e=JofXMr) )

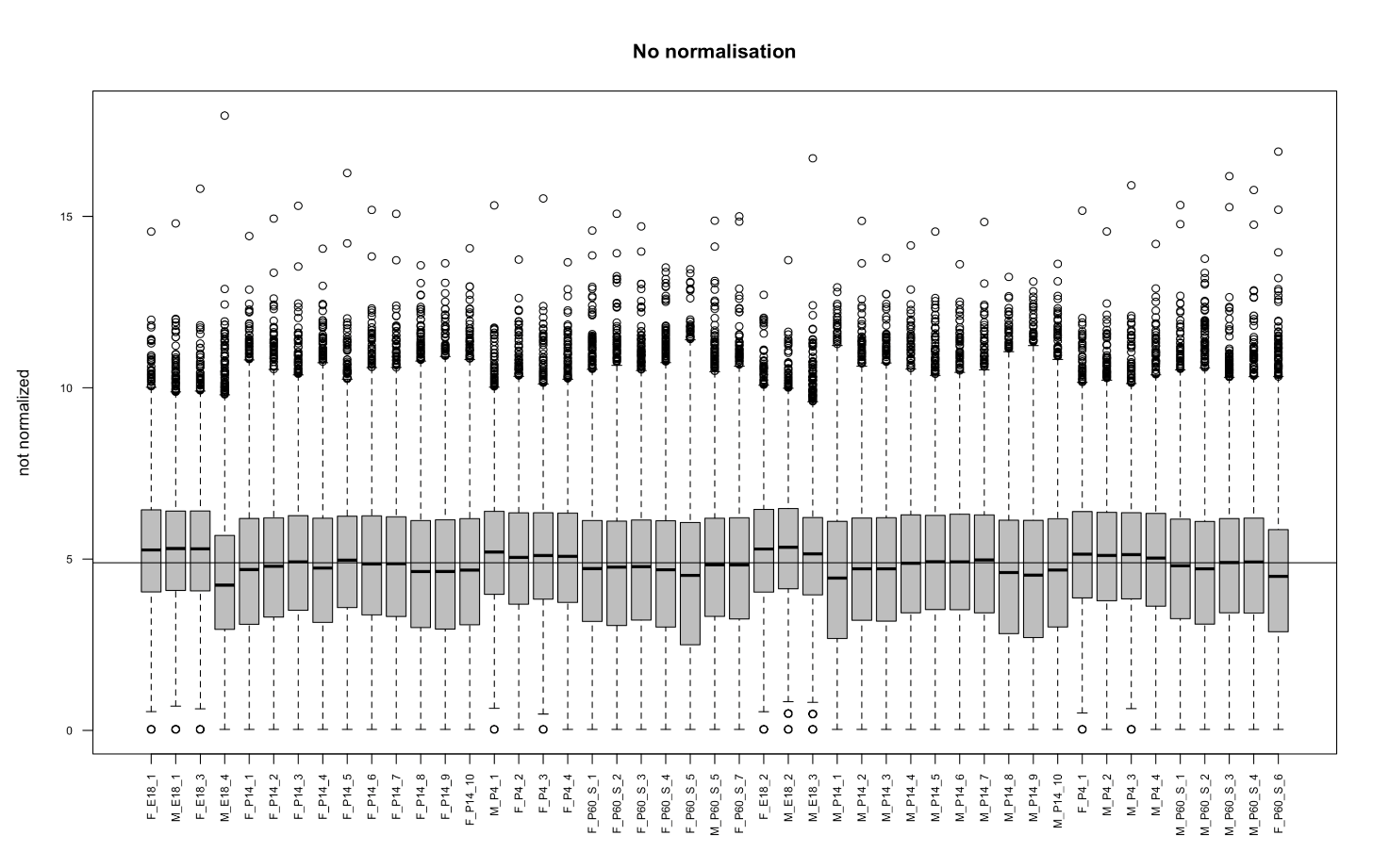


Fig 2: Box plot showing the non-normalized counts on y axis and samples on x axis

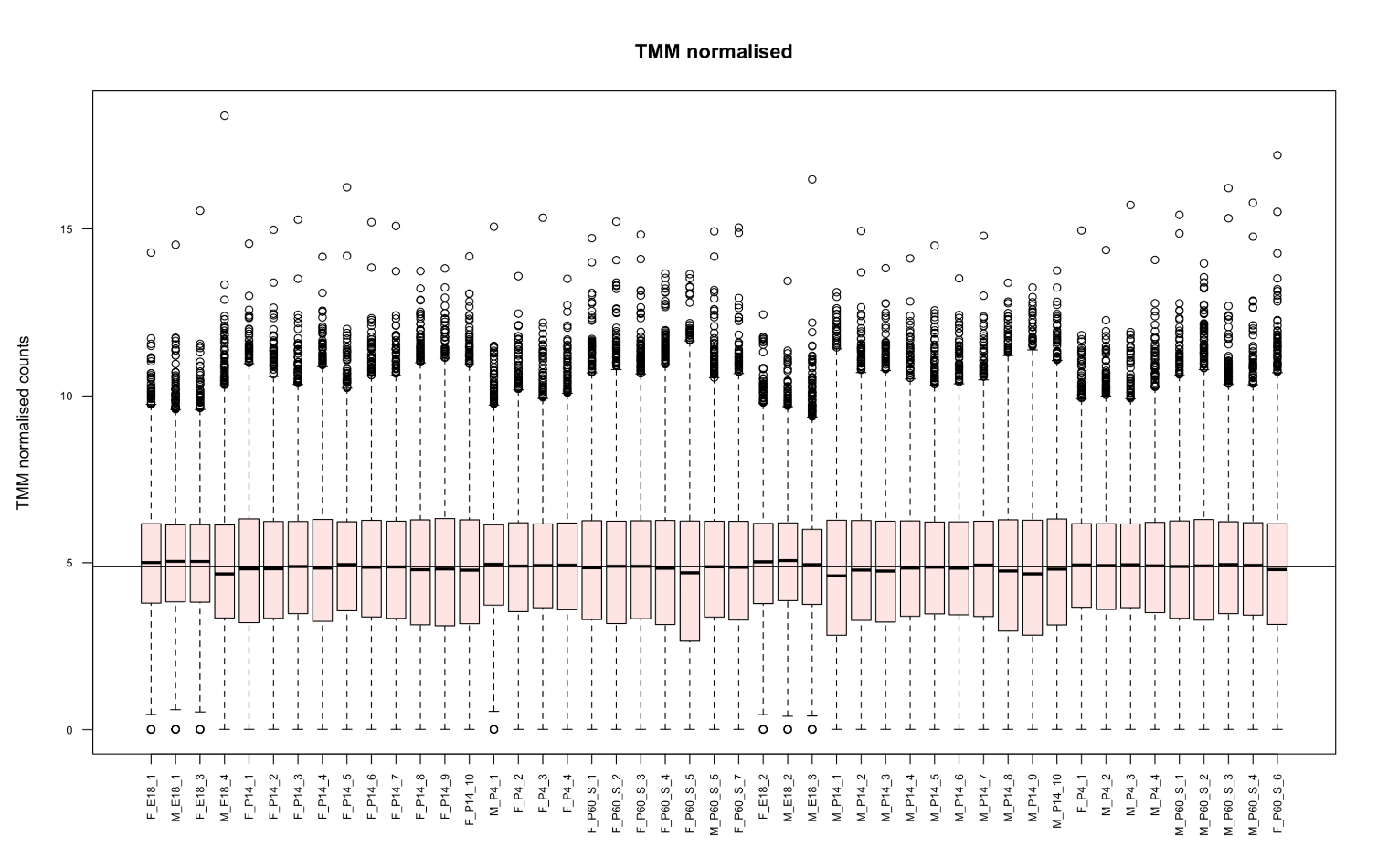


Fig 3: Box plot showing the TMM normalized counts on y axis and samples on x axis

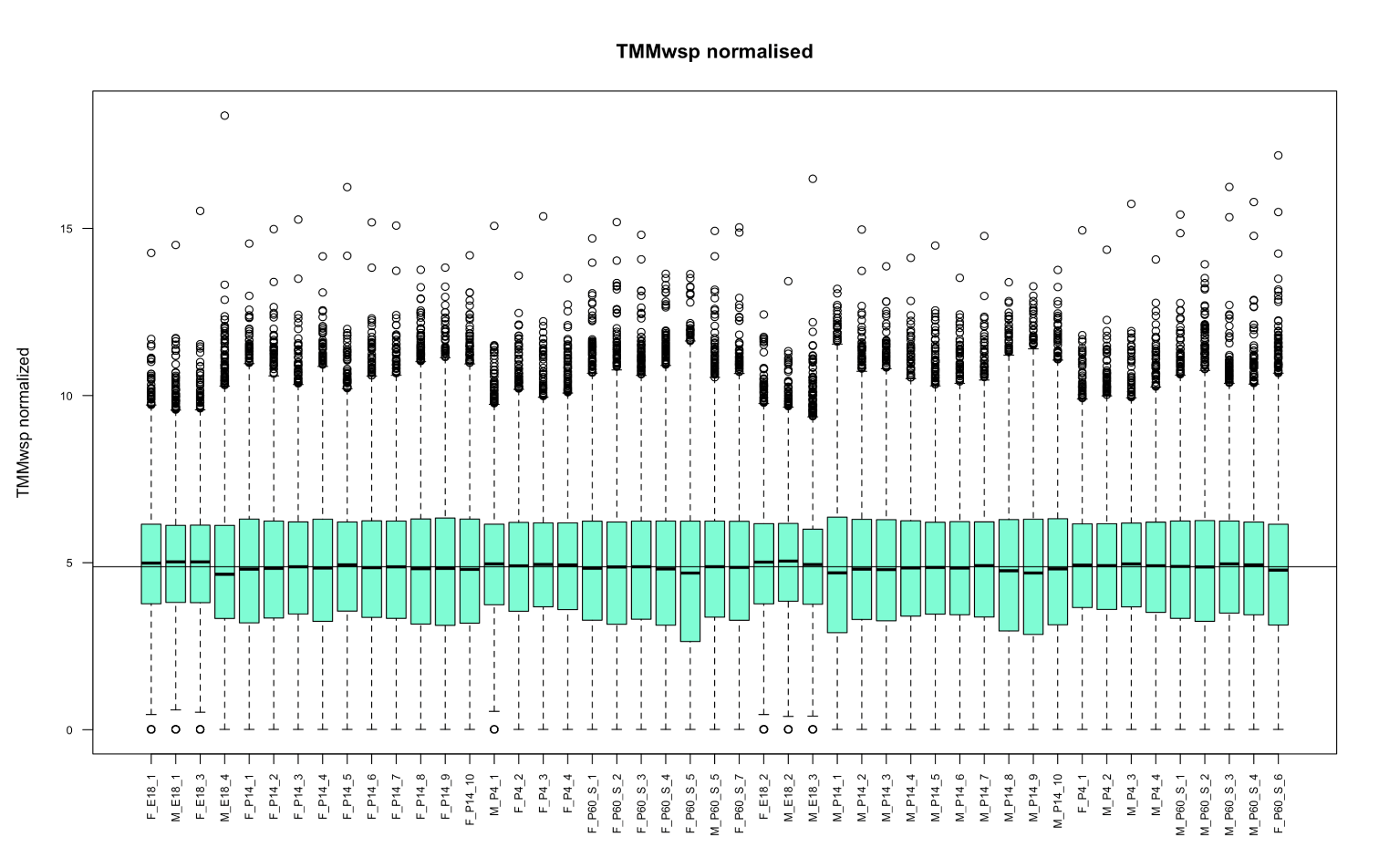


Fig 4: Box plot showing the TMMwsp normalized counts on y axis and samples on x axis

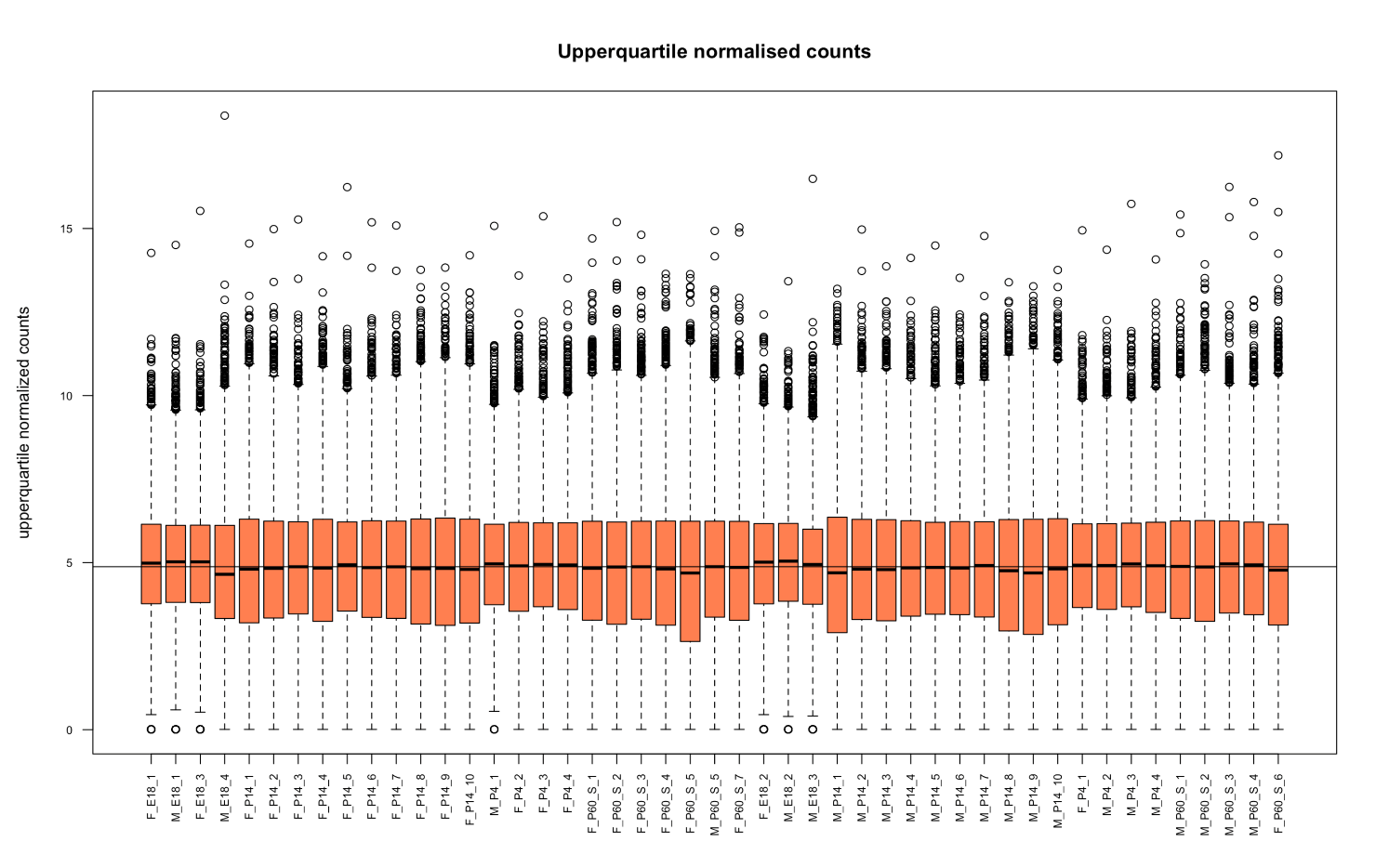


Fig 5: Box plot showing the UQ normalized counts on y axis and samples on x axis

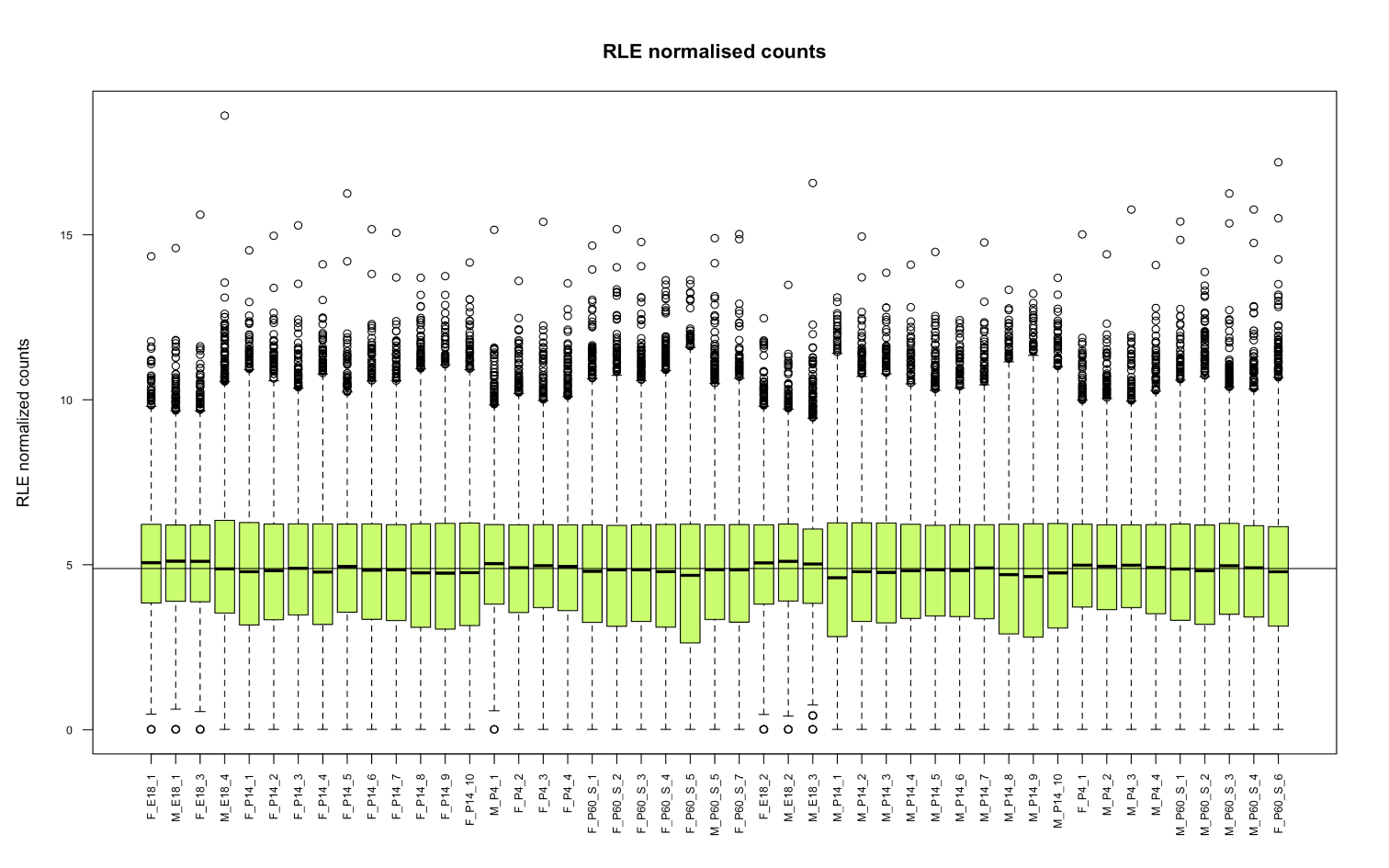
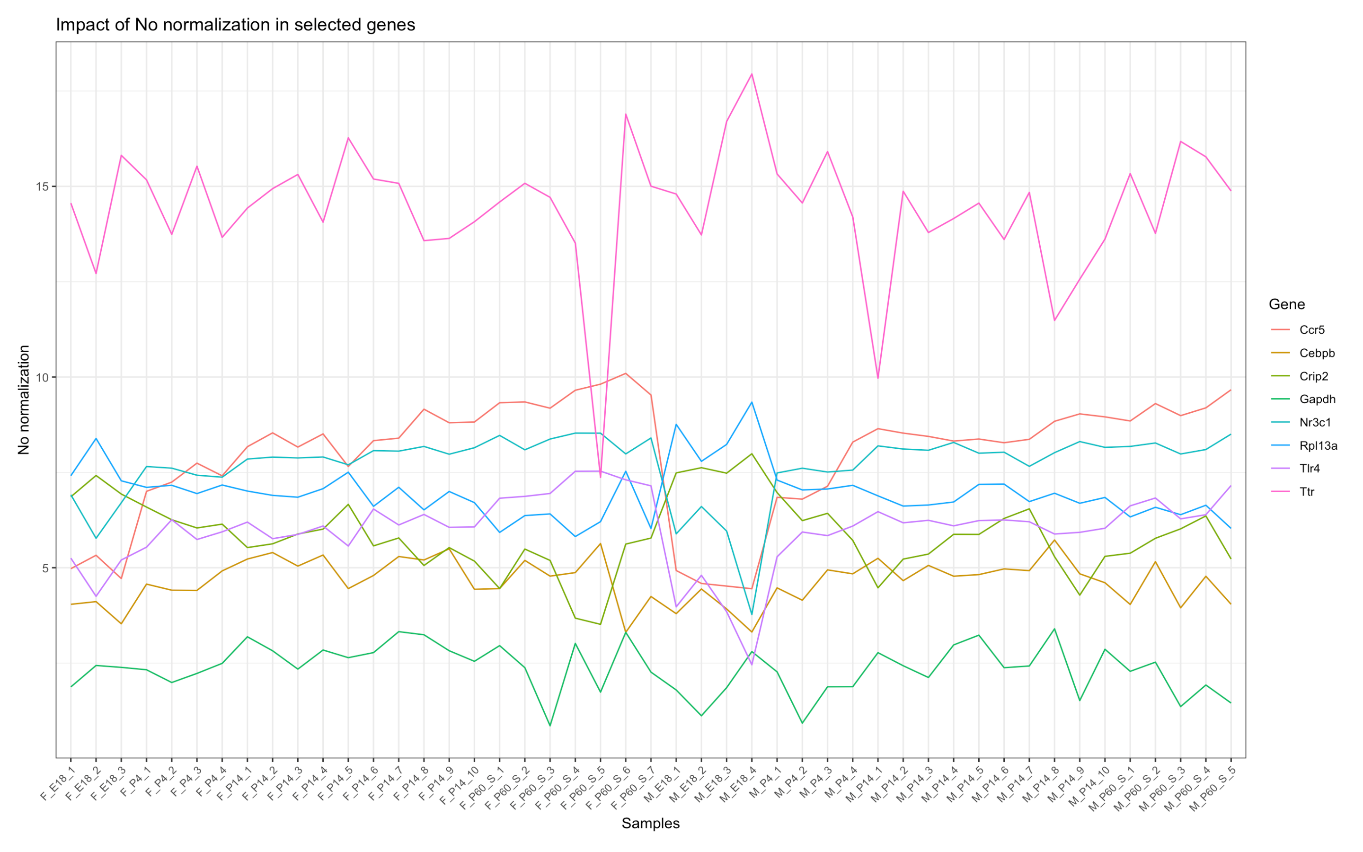
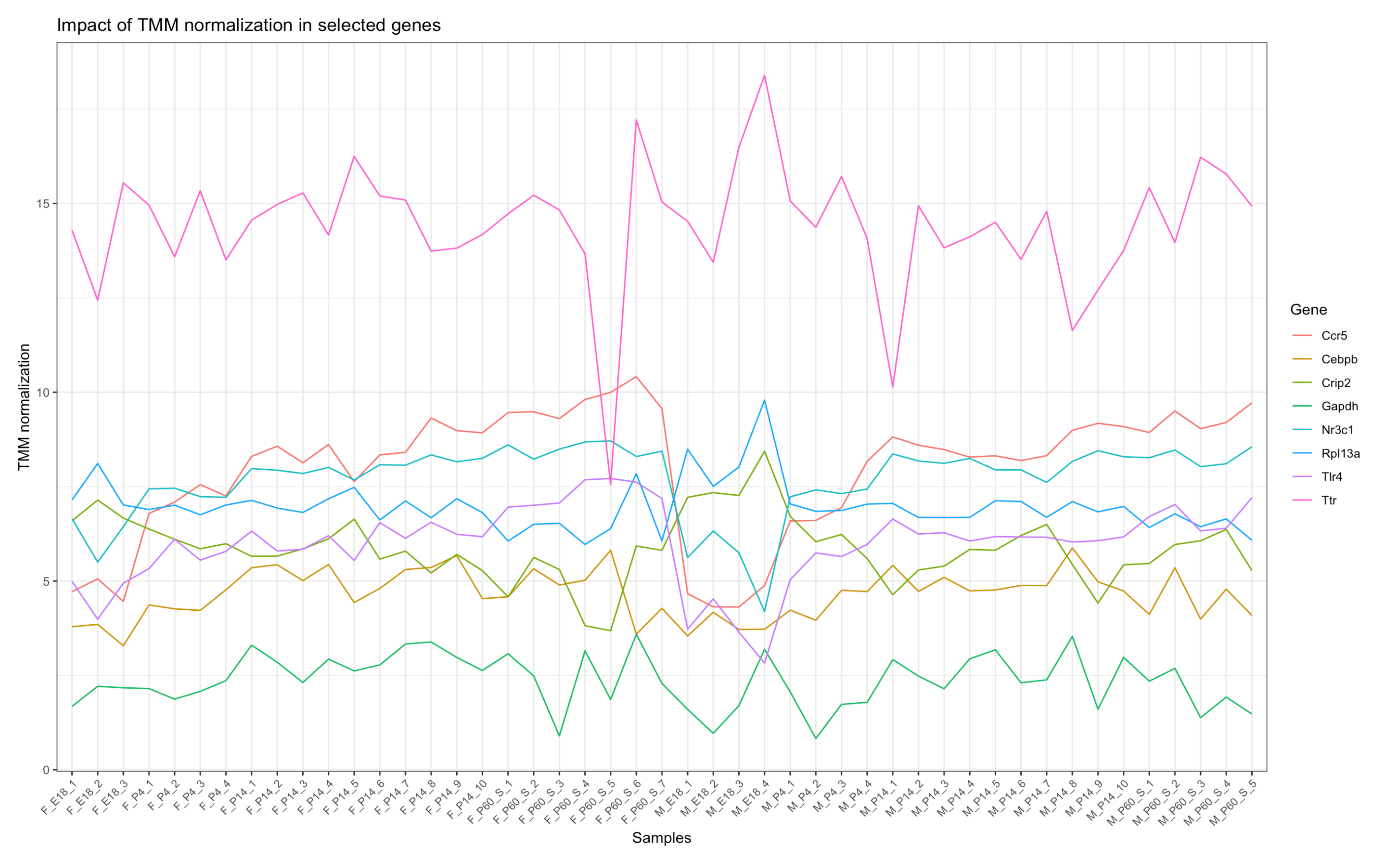


Fig 6: Box plot showing the RLE normalized counts on y axis and samples on x axis.

1. Impact of normalisation on individual genes

The impact of normalization on few selected genes were observed. We use Ttr gene, because it appeared to be compositional bias from the raw count distribution (Supplementary 3 [Distribution\_plot\_for\_raw\_counts.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Distribution_plot_for_raw_counts.html?csf=1&web=1&e=A7csby)) . We used control genes which are frequently used at lab for RT –qPCR Rpl13 and Gapdh. We also used our gene of interest Nr3c1 and Cebpb and immune related genes mentioned in the published paper of dataset Ccr5, Crip2, Gapdh, Rpl13a and Tlr4 (supplementary 3 [Impact\_on\_genes.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Impact_on_genes.html?csf=1&web=1&e=L8jeZ6)).





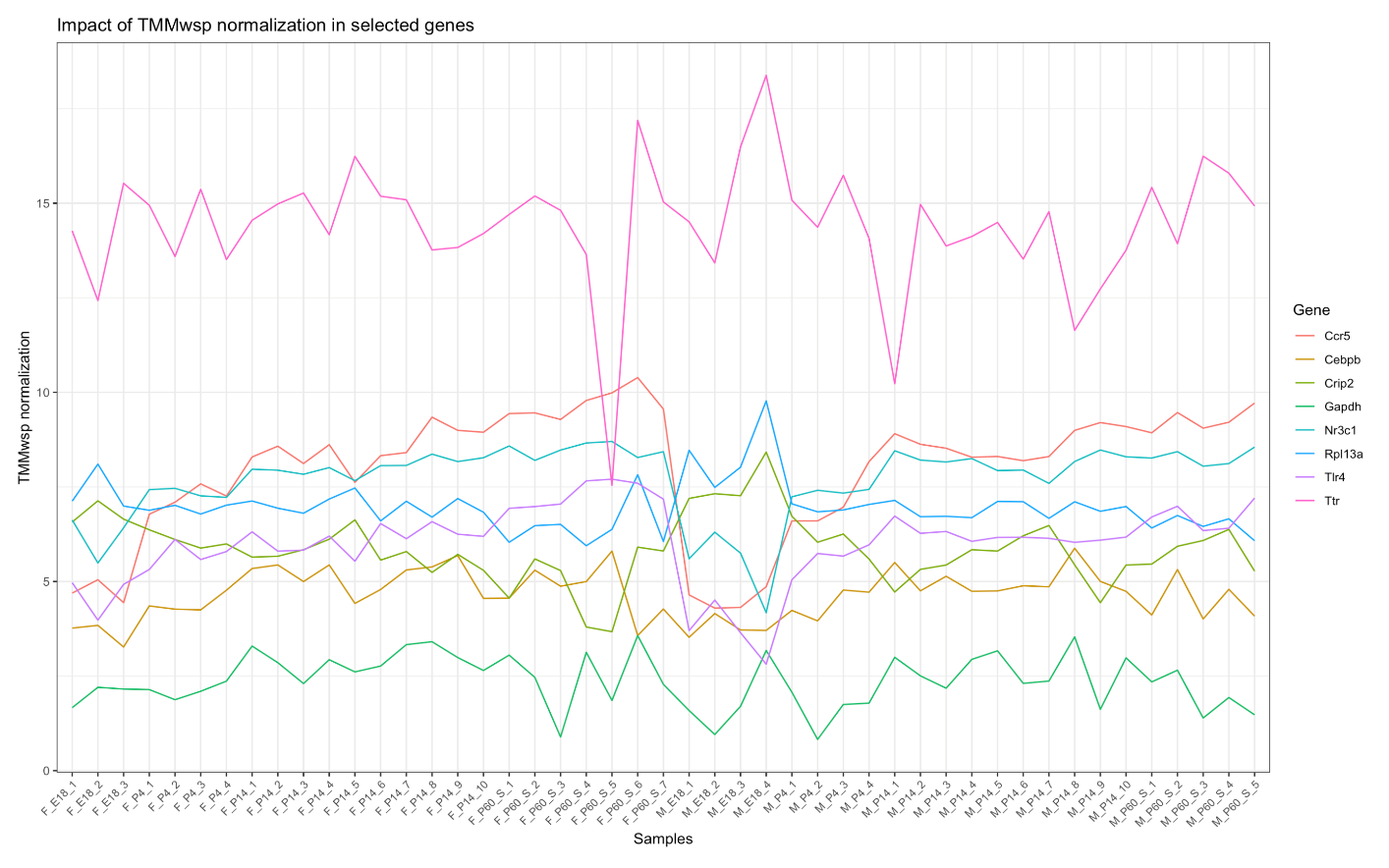


Fig: mmmmmmm

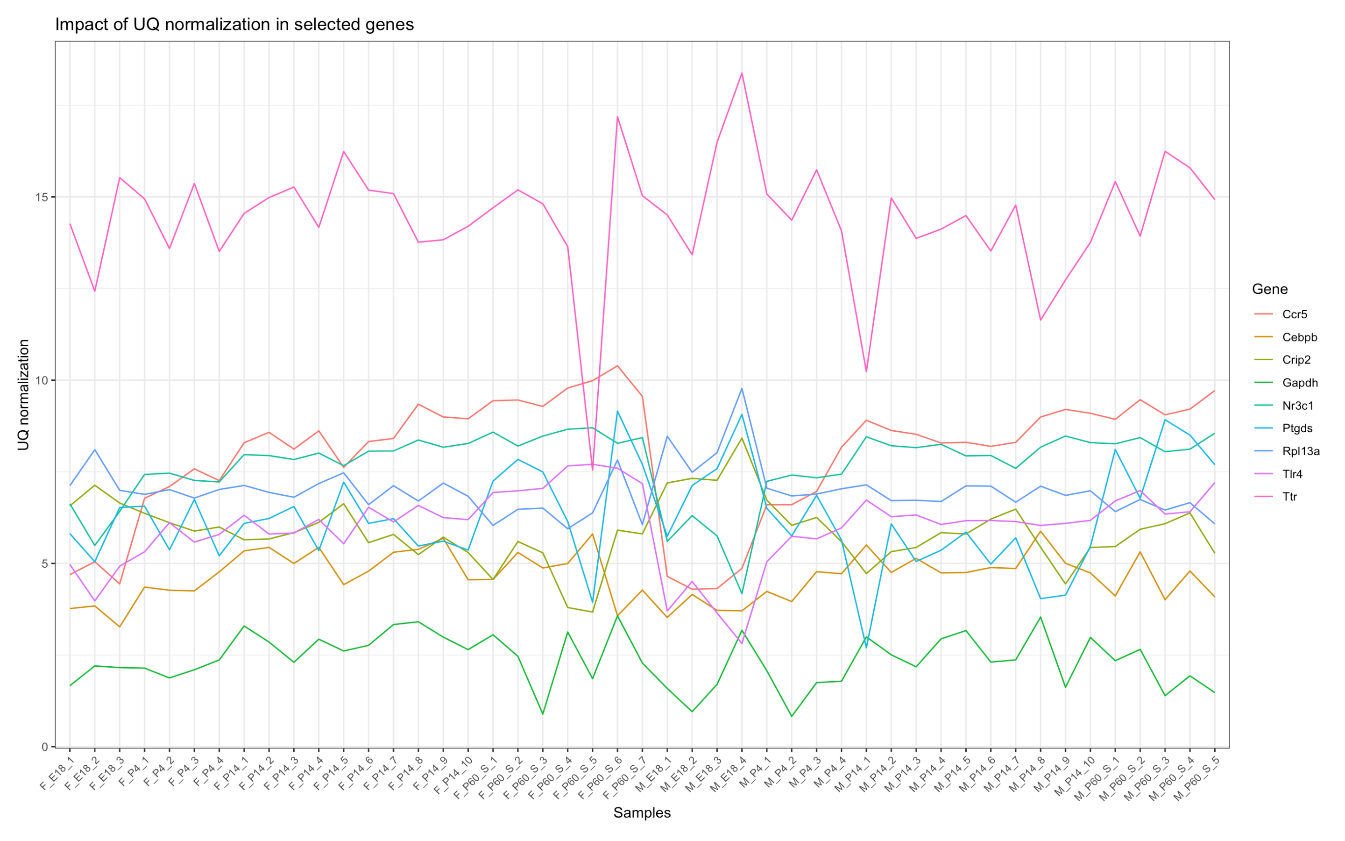
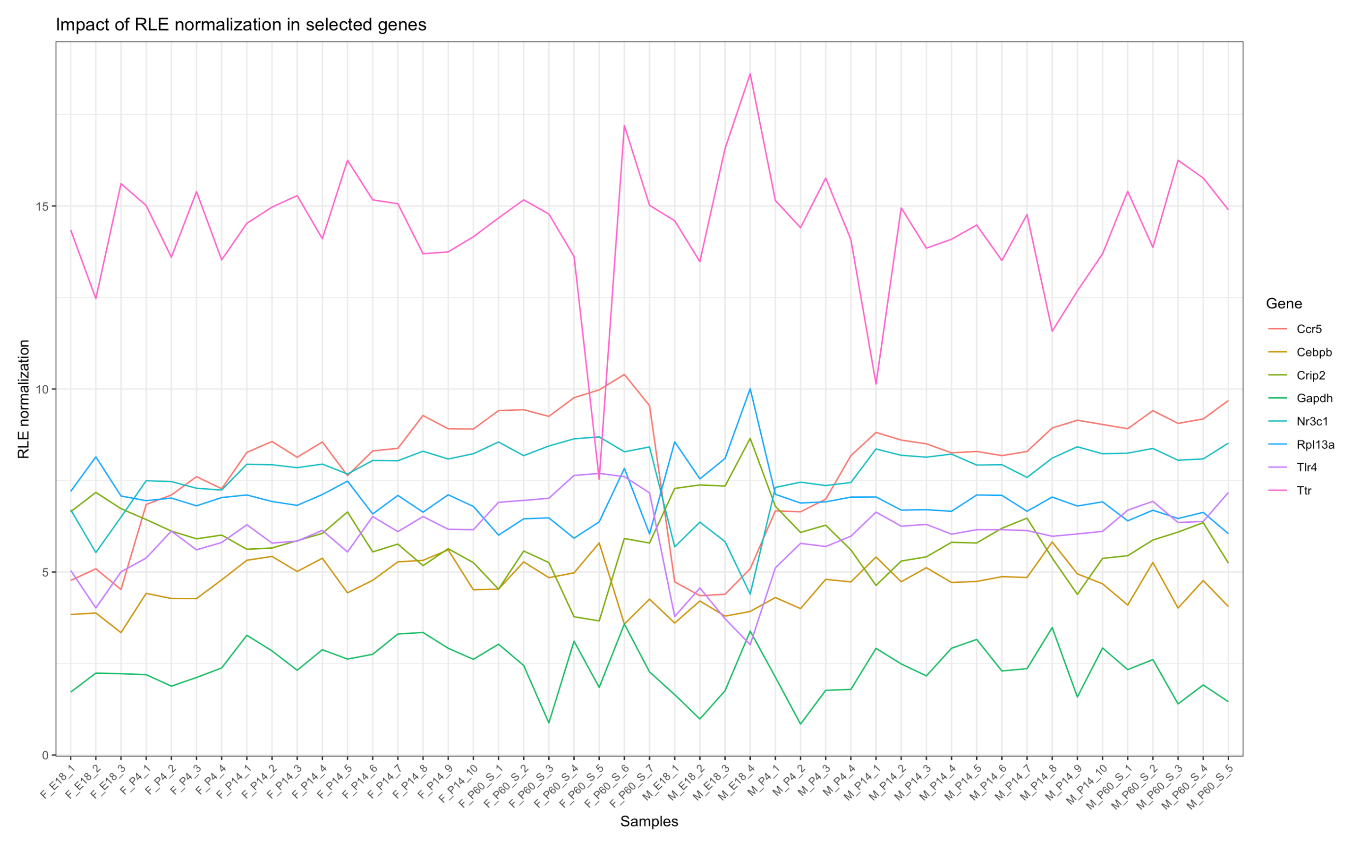


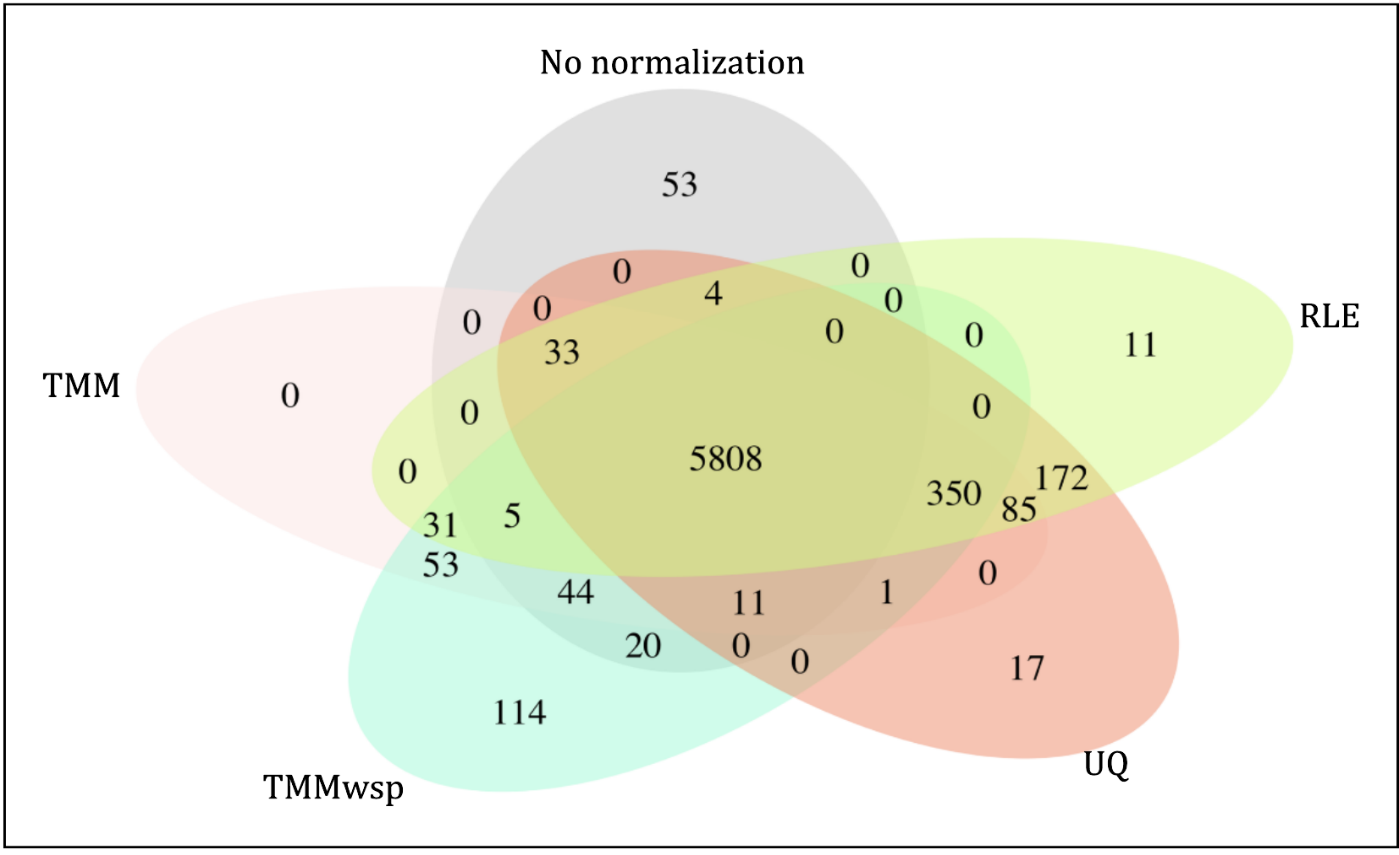
Fig: Impactlllllllllllllllllllllll

Fig : Impact of normalization

Slight difference was observed in differential gene expression analysis

Using edgR package differential gene expression between E18 and P4 was analysed.

The detail of differential gene expression is in Supplementary file 2.



The set of unique genes from above diagr

am is at (Supplementary file 4: [Unique\_elements\_of\_set.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Unique_elements_of_set.html?csf=1&web=1&e=rfczUn))

Batch effect correction

[Combat\_seq\_new.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Combat_seq_new.html?csf=1&web=1&e=N29wjn)

