**Normalization Strategies for Differential Gene Expression Analysis**

**Method**

**Dataset**

All normalization strategies were compared to each other by implementing same data set. The data was obtained from a paper “Generation of a microglial developmental index in mice and in humans reveals a sex difference in maturation and immune reactivity” published in Glia in 2017 and subsequent erratum at 2018 (Hanamsagar et al. 2018). The data was downloaded from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99622>.  In this analysis gene level differential analysis was performed and the LPS samples were excluded to look the developmental changes only in Microglial cells, however we will be extending our research for immune activated samples as well. The sample size is 47 with 3 replicates in Female embryonic 18 days (E18) and 4 replicates in male E18 days, 4 replicates for post-natal 5 day (P5) for both male and female, 10 replicates for P14 for both male and female and 7 P60 female replicates and 5 P60 male replicates, both treated with saline.

**In this study, four normalization techniques** TMM, TMMwsp, UQ and RLE were evaluated which are used for differential gene expression analysis. All four normalization strategies have an important statistical assumption that, the majority of genes are not differentially expressed. This assumption makes sure that the normalized counts of all the samples have similar distribution.

**Pre-processing:**

1. **Loading and Filtering of lowly expressed genes**

The raw data was retrieved from NCBI and subsequently saved in a directory. The raw data was made read only so that it would not be manipulated during the data munging process.

# Make data immutable from terminal

% chmod 400 GSE99622\_hanamsagar2017\_raw\_reads.txt

The data was loaded with genes as row names and samples as column names. edgR function *filterByExpr* was used to filter lowly expressed genes*.* This function keeps the rows with worthwhile counts. This filtering keeps genes that have CPM above “k” in “n” samples, here “k” minimum count according to the library size and “n” is the smallest group size (Chen, Lun & Smyth 2016).

keep <- filterByExpr (y, design)

y <- y[keep,]

1. **Choosing negative control**

Negative control is CPM normalized counts which is log transformed. All other methods of normalization were used after doing CPM normalization which adjust the sequencing depth bias and are log transformed. Method is “none” for no normalization that gave negative control. “TMM” for TMM normalization, “TMMwsp” for TMMwsp normalization, “RLE” for RLE normalization and “upperquartile” for UQ normalization.

no\_normalise <- calcNormFactors(y, method = "none")

logcount <- cpm(no\_normalise,log=TRUE)

1. **Exploratory data analysis**

The histogram and box plots of all the samples using all four methods and negative control were created to know the distribution of the sample. For the histogram gaussian filter was used to easy visualization.

1. **Getting the differentially expressed genes**

The differentially expressed genes (DEGs) using all four methods and negative control were obtained using *edgR*. Here for example, DEGs from UQ is being calculated.

UQ <-estimateDisp(UQ,design, robust = TRUE)

UQ <-glmQLFit(UQ,design, robust =TRUE)

UQ <-glmQLFTest(UQ, contrast = E18vsP4)

The *estimateDisp* function maximizes the negative binomial likelihood to give the estimation of common, trended, gene wise dispersion across all the genes, then empirical Bayes method was used to estimate posterior dispersion, robust = TRUE means, the estimate should be robust against the outliers. *glmQLFFit* function performs empirical Bayes quasi likelihood F-Tests. The contrast matrix was set E18 – P4, as an example, other matrix of interest can also be chosen. Multiple test correction was performed by applying the Benjamini-Hochberg method on the P values to control False positive discovery rate (FDR).

1. **Testing the Batch effect correction on the data**

Likelihood of presence of batch effect on large dataset such as one being analysed here is high. Since there were replicates ranging from 1 to 10 in different samples, ComBat seq in *SVAseq* package was used. ComBat seq uses empirical Bayes method to estimate parameters to adjust the batches and output are positive integers making easy to use for downstream analysis.

The MDS (Multidimensional scaling) plot was used to explore difference between the batch corrected and uncorrected data. MDS clusters the similar samples together.

**Results**

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

Significant differences in the distribution of the samples were not observed using different normalization techniques as an example the histogram for F\_E18\_1 (Female embryonic day 18 replicate 1). The distribution was near to Normal distribution, hence, for downstream differential analysis parametric tests would be suitable. The peak of curve was at around 5 where is mean, median and mode lies. (Code with all the histograms of all the samples are in (code and analysis: [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 graph of a normalized number

Description automatically generatedA graph of a normalized number with Ryugyong Hotel in the background

Description automatically generated

B

A

A green and blue graph

Description automatically generated

D

C

A graph of a normalized number

Description automatically generated

E

A green and black line graph

Description automatically generated

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 (negative control) B) TMM normalization method C) TMMwsp normalization method D) UQ normalization E) RLE normalization

1. **Only negative control data have fluctuation of median and quartile values than other normalized data**

All the samples had similar median values observed in the boxplots in all four normalisation approaches except for the boxplot negative control where slight fluctuation was observed. This showed that four approaches are making the samples more comparable than no normalization. The abline is created from median value to entire dataset. (Code and analysis : [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) )

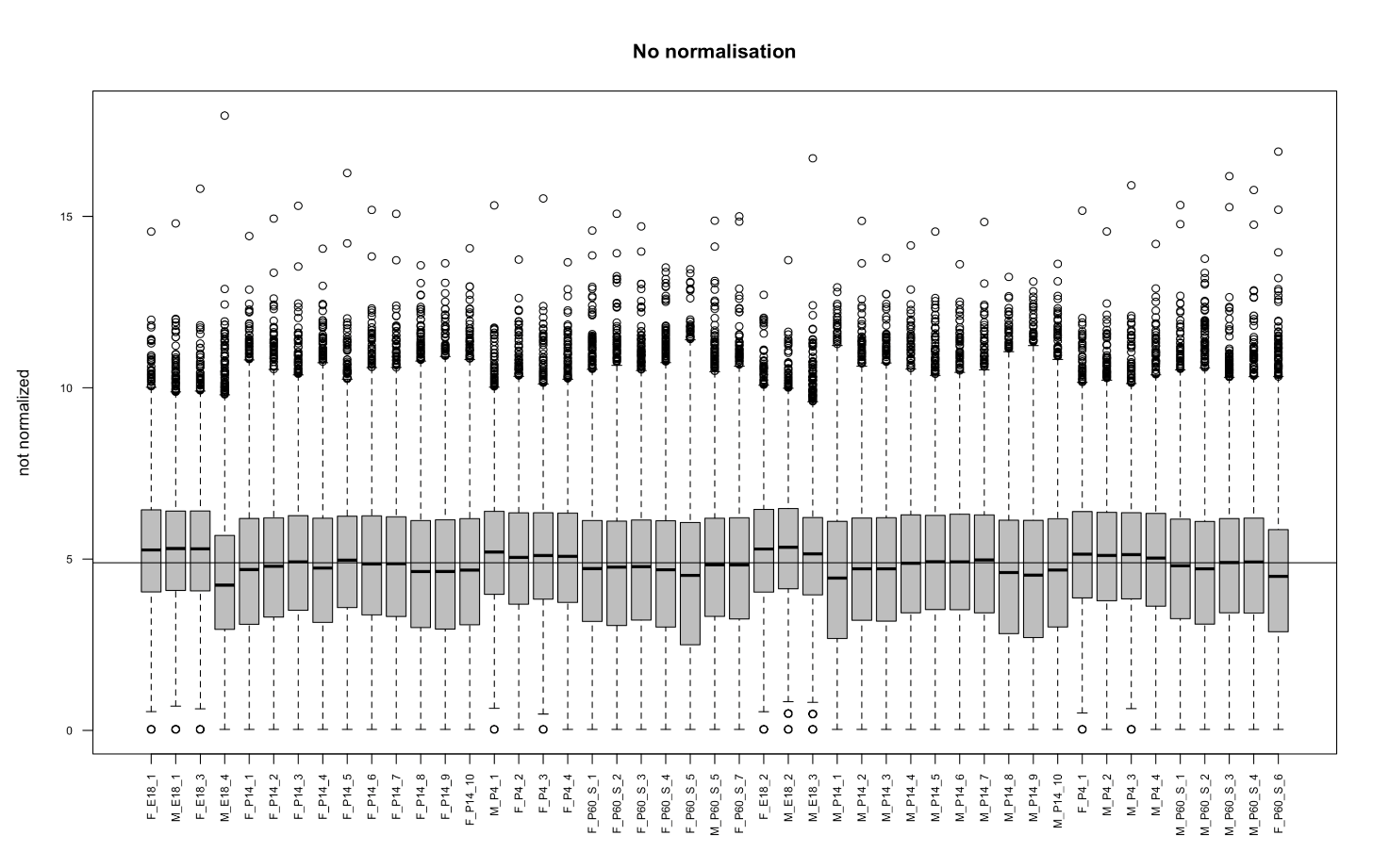


Fig 2: Box plot showing the non-normalized counts on y axis and samples on x axis. Median values are slightly fluctuating from abline.

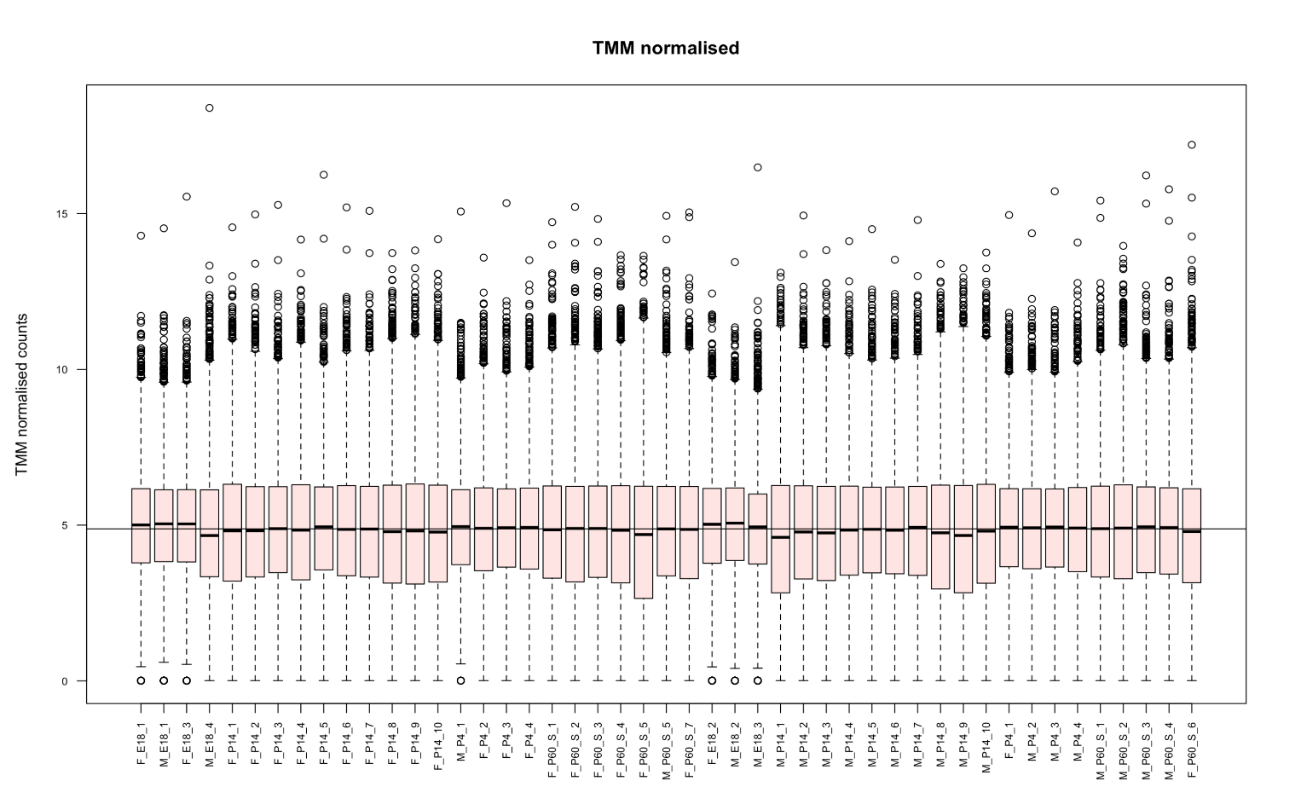


Fig 3: Box plot showing the TMM normalized counts on y axis and samples on x axis. Median values are not so much fluctuating from abline.

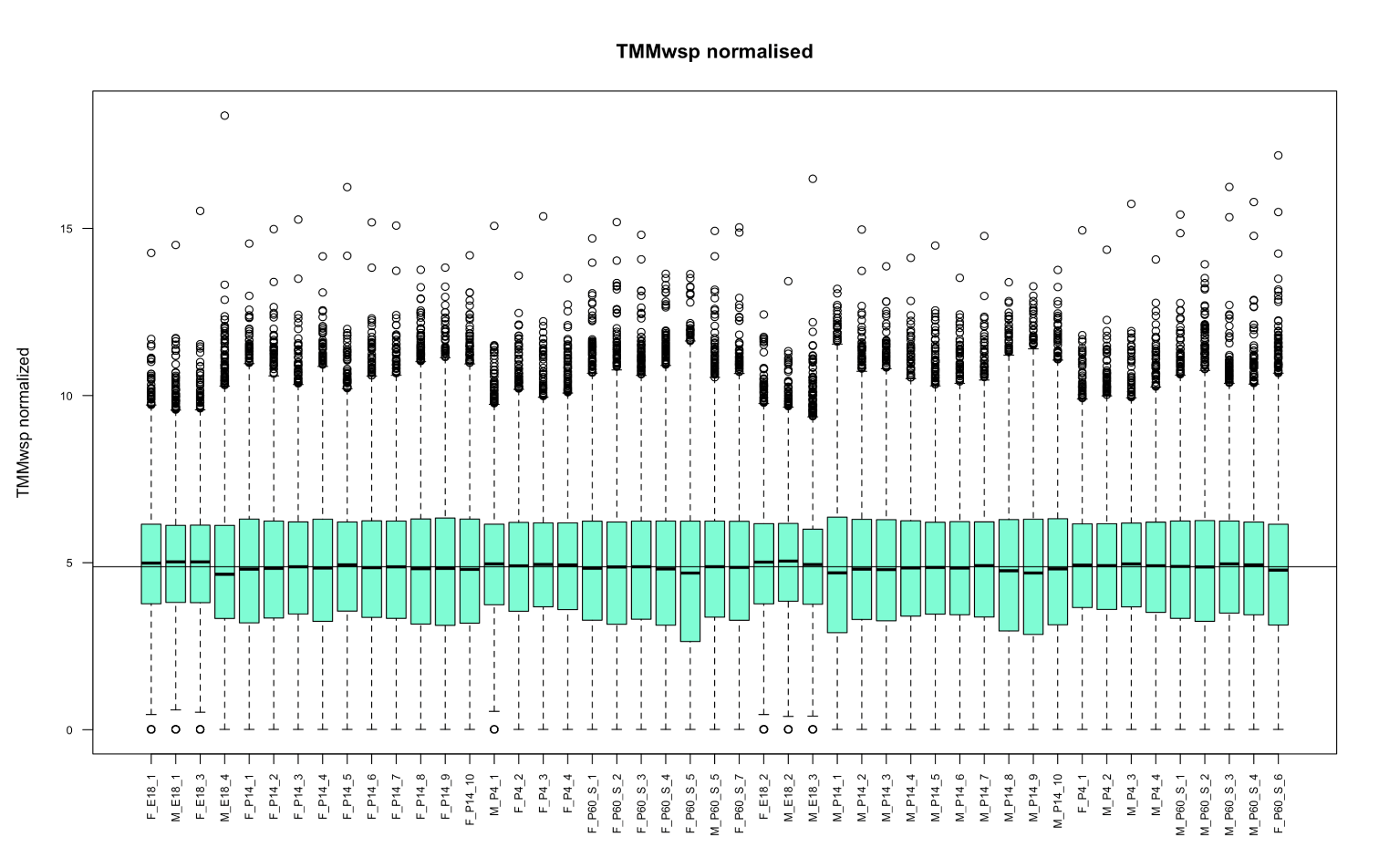


Fig 4: Box plot showing the TMMwsp normalized counts on y axis and samples on x axis. Median values are not so much fluctuating from abline.

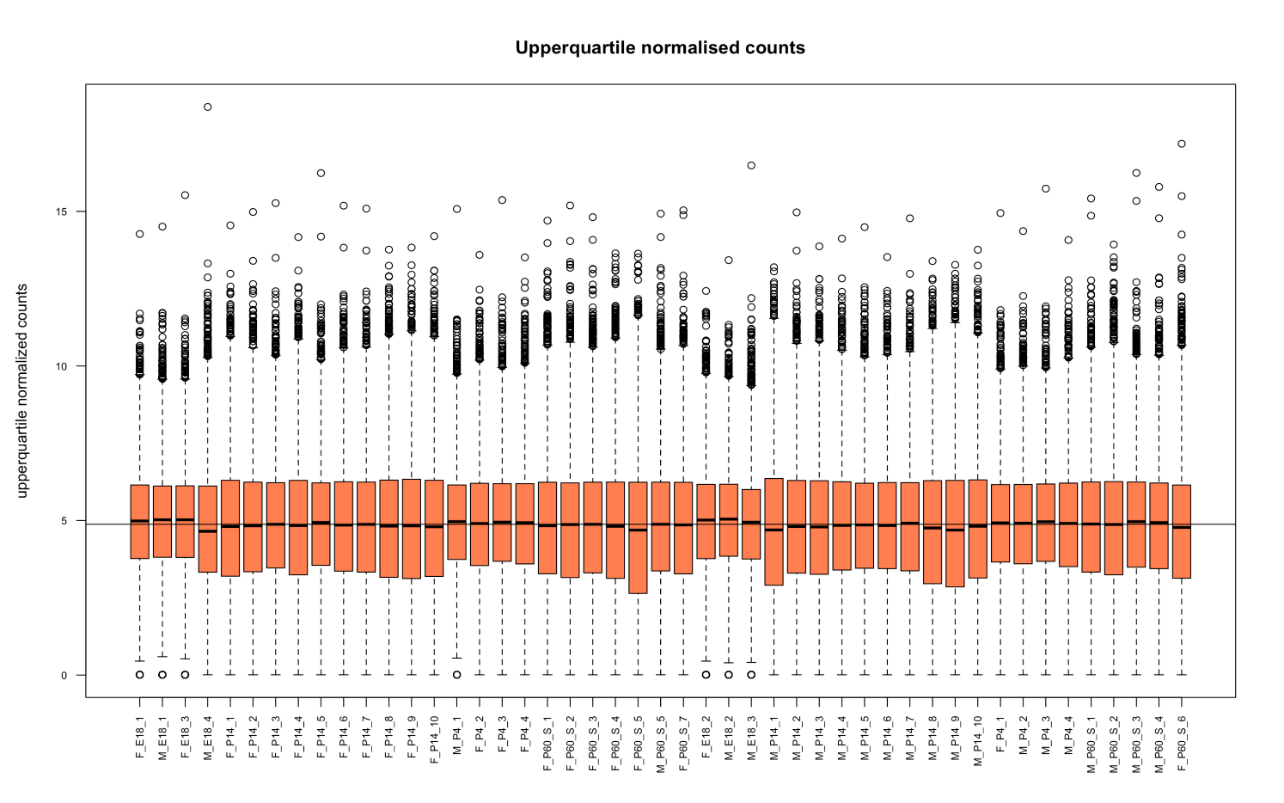


Fig 5: Box plot showing the UQ normalized counts on y axis and samples on x axis. Median values are not so much fluctuating from abline.

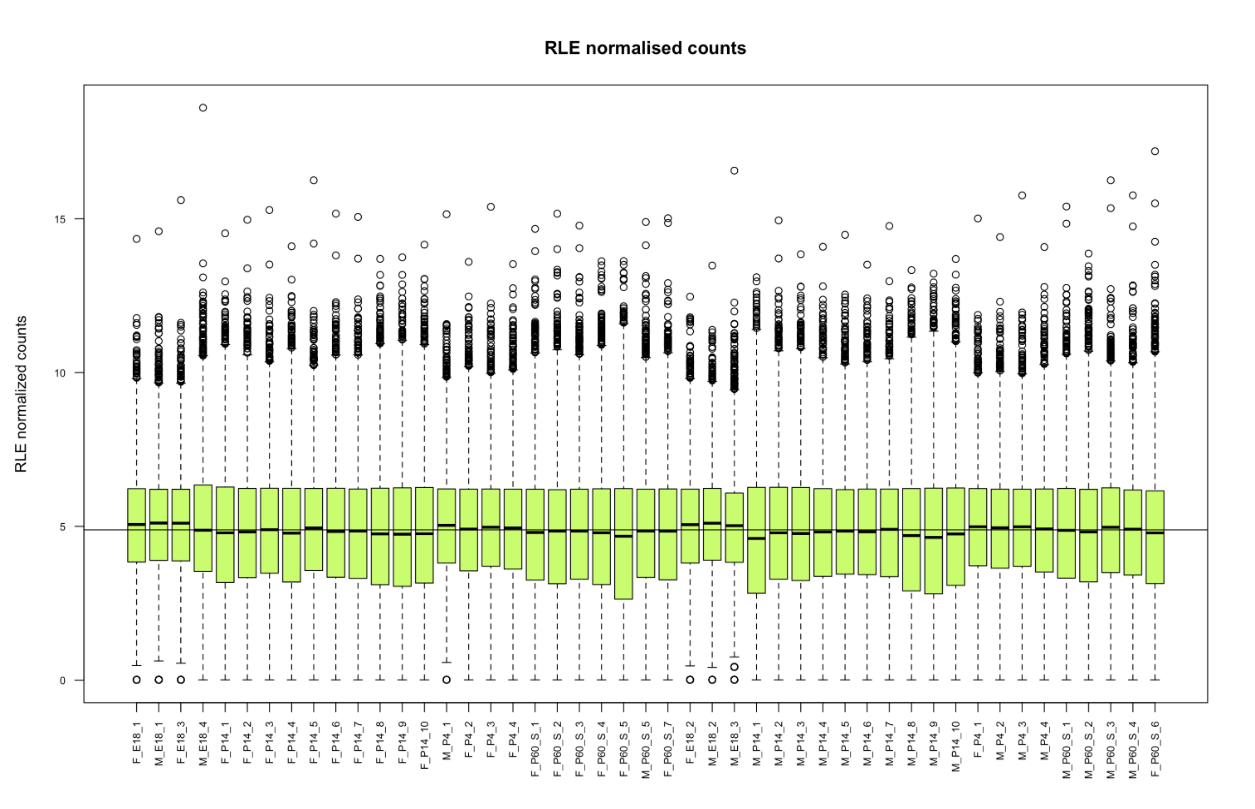


Fig 6: Box plot showing the RLE normalized counts on y axis and samples on x axis Median values are not so much fluctuating from abline.

1. **Difference was observed in differential gene expression analysis**

Differentially expressed genes (DEGs) were obtained using *edgR* package and contrast matrix was set E18 - P4, this contrast as an example. Highest number of DEGs, 6499, were obtained from RLE normalization, which was 10.63% higher than common genes between all the techniques, followed by UQ. RLE and UQ are the techniques which does not trim the extreme values like TMM and TMMwsp does. The lowest number of DEGs were obtained from negative control, only 5978 total DEGs. Net 114 unique DEGs were obtained from TMMwsp which is predictable result as its algorithm includes the genes which have at least one positive count while comparing for pair of libraries, whereas TMM ignores gene if it encounters the zero count while comparing a pair of libraries. This also explains why the TMM has no unique set of genes. The detail of differential gene expression is in Supplementary file 2 and set of unique genes along with fold change and P values are in Supplementary file 3. ( [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))

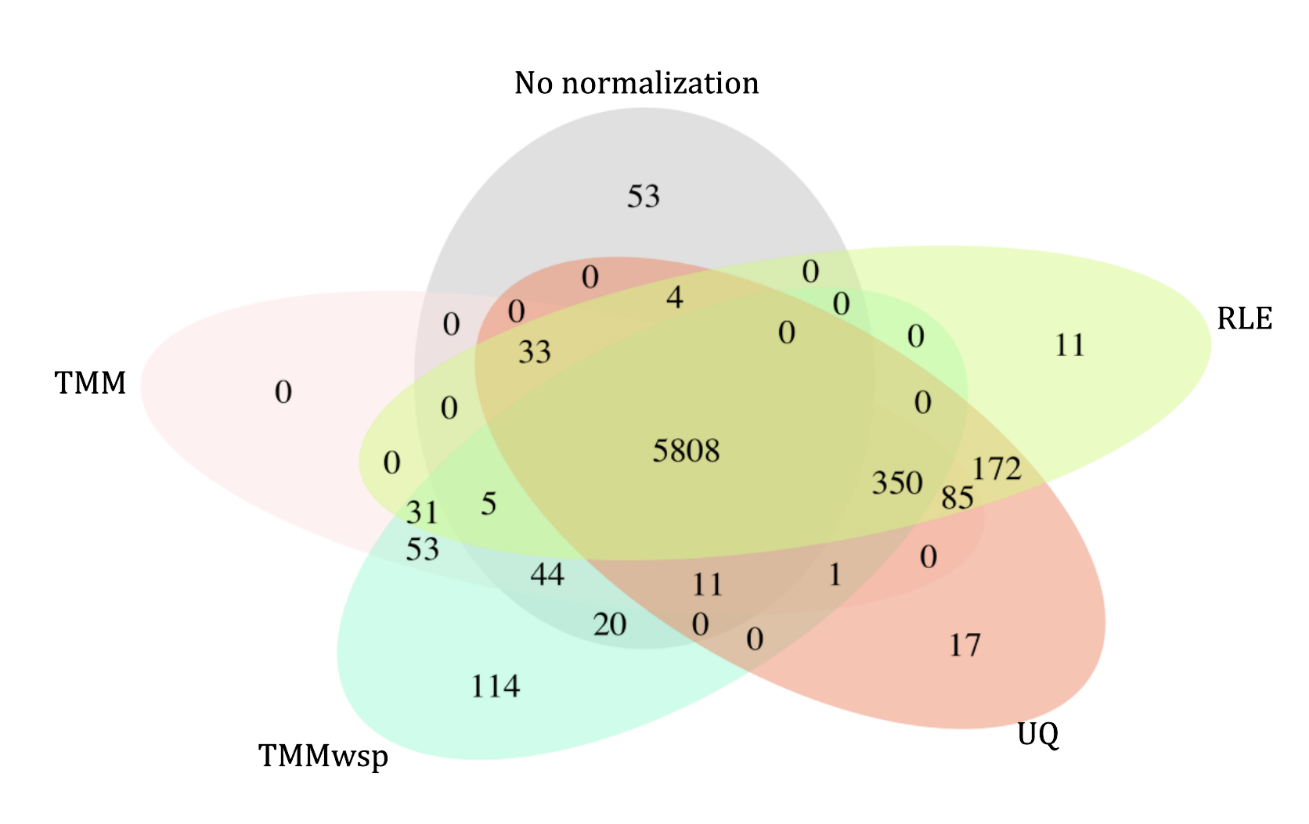


Fig 7: Venn diagram of DEGs using various normalization strategies. 5808 genes are overlapping between all methods and control. Total number of DEGs detected are 5978,6421, 6437, 6481, 6499 respectively from no normalization, TMM, TMMwsp, UQ and RLE techniques.

1. **Batch effect correction with ComBat seq did not produce meaning result**

From MDS plot after correcting for the batch effect by ComBat seq, variance among the sample was significantly decreased, approximately by 21% and clustering of P4 samples with P14 samples was observed which is biologically unreasonable. Therefore, the differential gene expression was performed without batch correction of this dataset. Code for batch effect correction in [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)

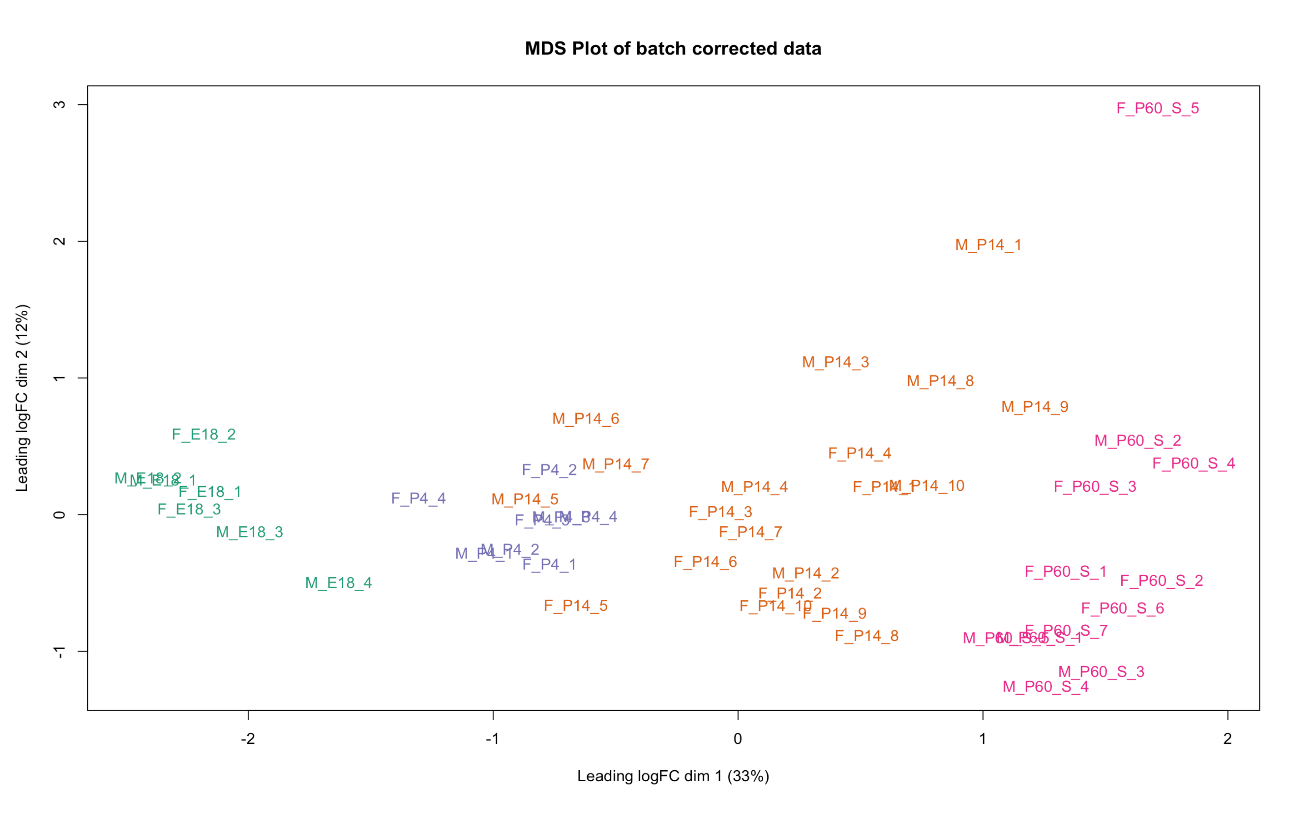


Fig 8: MDS plot of batch corrected data. The x axis isre presenting dimension 1 of batch corrected values which is CPM normalized and log transformed, and a y axis is presenting dimension 2. Colouring by age group. Variance in dimension 1 is 33% and P4 samples are clustered with P14.

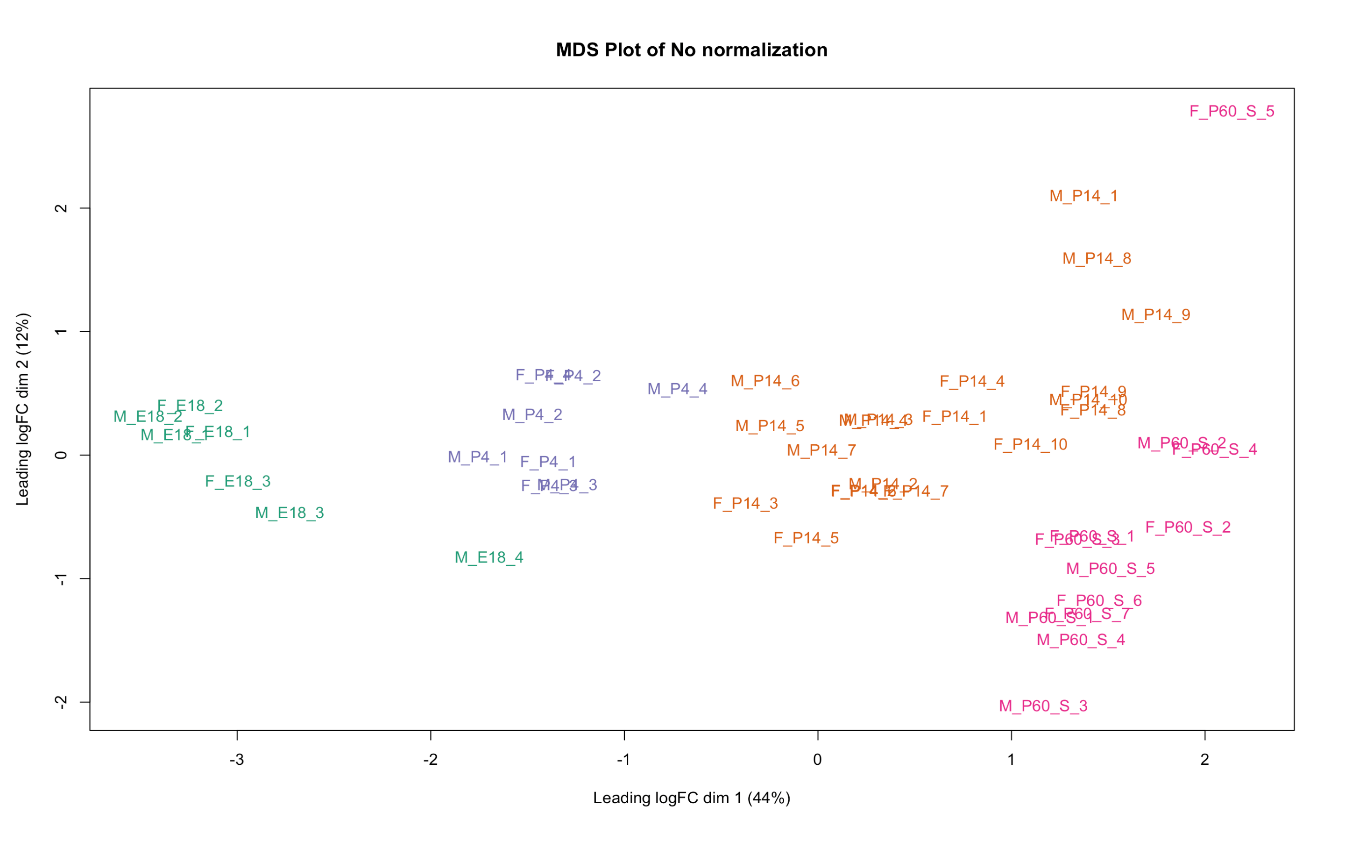


Fig 9: MDS plot of data without batch effect correction. The x axis is representing dimension 1 of values which is CPM normalized and log transformed, and a y axis is presenting dimension 2. Colouring by age group. Variance in dimension 1 is 44% and samples are clustered according to age on dim 1.

**Conclusion**

Since, majority of differentially expressed genes are overlapping, any one of the four techniques can be used for this dataset. The best way to determine which technique is the most precise and accurate showing the true biological signal would be conducting the RT -qPCR for the unique gene sets we have from the Venn diagram. From the statistical point of view, TMMwsp has a more inclusive algorithm, it does not remove genes while comparing the two libraries even if one of them has zero count, unlike TMM which does remove. I think TMMwsp can be used as method of preference to minimize gene loss due to zeros since the dataset we are using has inflated zeros and also it can handle extreme outliers. However, for validation of this *in silico* conclusion, in *vitro* or *in vivo* experiments must be conducted.

**References**

Chen, Y, Lun, A & Smyth, G 2016, 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]', *F1000Research*, vol. 5, no. 1438.

Hanamsagar, R, Alter, MD, Block, CS, Sullivan, H, Bolton, JL & Bilbo, SD 2018, 'Generation of a microglial developmental index in mice and in humans reveals a sex difference in maturation and immune reactivity', *Glia*, vol. 66, no. 2, pp. 460-460.