Problem 1

Two group comparison: Apply EBSeq to the PBMC data to determine the number of genes differentially expressed (DE) between the two conditions.

1. How many genes are DE at false discovery rate (FDR) 5% and 10%? (using maxround=5 of iterations)
   1. 3001 at 5%
   2. 3396 at 10%
2. What are the estimated library size factors? What information do they provide? (3 pts)

All are the same value: 1.006661

They represent the relative size of each experiment's total number of sequencer reads. Comparing the values returned by MedianNorm() gives relative library sizes of each column in your df.

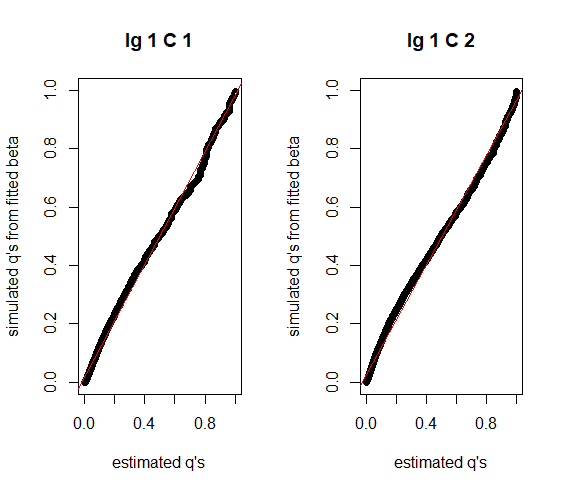
I’m unsure why the values returned by MedianNorm are not themselves centered to 1. It feels like they should be.

1. Does the model fit well as assessed via diagnostics? Please show diagnostic plots and discuss. (3 pts)

QQ Plot:

The data points lie near to the y=x line, which suggests that using a Beta prior is appropriate.

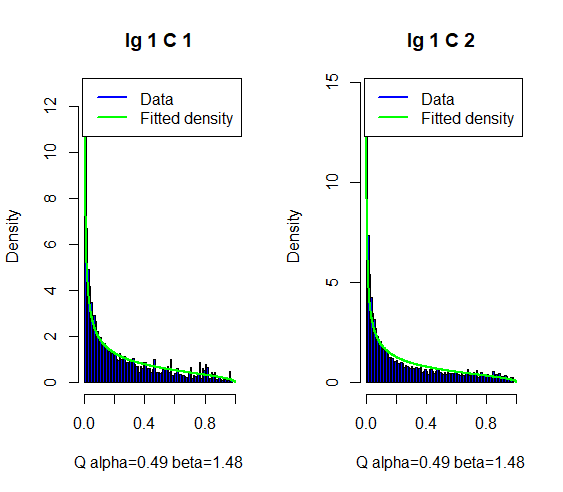
There is a mild deviation on Condition 1 plot near 0.8, but it is doubtful this presents any meaningful problem.



The fitted density closely matches the empirical distribution, suggesting that our assumptions hold, and the distribution from our data matches the model.

If there is a large deviation from the fitted line, what does that say about our experiment? Bad experimental design? Bad sequencing run?

What do real examples of poorly fitted empirical distributions look like?

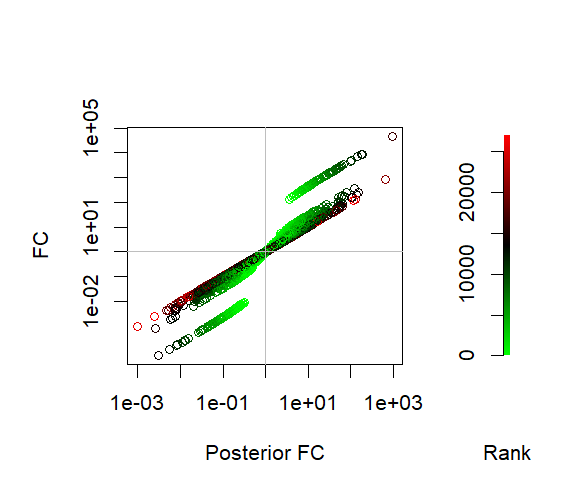


I don’t know if this counts as a diagnostic plot, but it’s still interesting.

There are two groups that deviate from y=x, and they have relatively high ranks (given by green color).

The tutorial vignette states that the posterior FC tends to shrink genes with low expressions (small rank).

But we see that relatively high ranked genes are shifted. I don't use R, so I can't dig into why that is and which genes are deviating. What’s going on here?



4-group differential expression analysis

1. Number of genes significant

Iterations were taking 15 minutes apiece, so I did only 1 iteration (maxround=1)

* 1. 5% FDR: 12952
  2. 10% FDR: 13323

1. Estimated library size factors are different from the PBMC data. the a549 data are approximately size 1.15. The NHBE data are about 0.89.

This suggests that the NHBE data had lower library counts than a549. It shows that normalization is important in this case to get accurate comparisons because there is a large difference between the two datasets.

1. In the study, they refer to the untreated, control cells as mock, which strikes me as an odd word choice. The genes that differ between each mock are those that naturally differ between each cell type. Different cell types express different levels of genes. This is a baseline level of differential expression in genes.
2. To compare control vs treatment, compare within each cell type, i.e. A549s mock vs treatment, and NHBE mock vs treatment. If your group ordering is:

A549s Covid A549s Mock NHBE Covid NHBE Mock

Then the DE pattern of interest might be 1 2 1 2.

However, this pattern might be misleading, because of the point in question #3 above. It could be that the gene has a different level of expression in the different cell types. And it could be that each cell has a different gene level response to the virus. In this case, then you would expect to see 1 2 3 4 as the pattern.

Therefore, the 1 2 1 2 pattern covers the case where gene expression baselines are identical between cell types naturally, and their gene level response after SARS-CoV-2 is identical. This is a restrictive case, and because of that I would be tempted to perform a separate 2-group comparison for each cell type and then assess DE.