quantro\_analysis

Daniel Montemayor

April 13, 2018

# A Quantro Analysis for Determining the Appropriateness of Quantile Normalization.

## Summary

Global differences in the distribution of the data exists. Quantile normalization will wash out this information and is therefore **not** appropriate.

## Setup requirements

## install rmarkdown  
if(!require("rmarkdown")){install.packages("rmarkdown")}

## Loading required package: rmarkdown

## Warning: package 'rmarkdown' was built under R version 3.4.4

library(rmarkdown)

First, Install quantro. Press a for “all” if prompted to update all/some/none packages.

## try http:// if https:// URLs are not supported  
if(!require("quantro")){  
 source("https://bioconductor.org/biocLite.R")  
 biocLite("quantro")  
}

## Loading required package: quantro

## Setting options('download.file.method.GEOquery'='auto')

## Setting options('GEOquery.inmemory.gpl'=FALSE)

## Data Cleaning

Next, read in your data. Should be samples in rows and observations in columns. We expect the first 2 columns to be sample number and sample group. The reamaining columns should contain the measured metabolite concentrations. You will need to name this file ‘quantro\_analysis.in.csv’ and save it in the working directory.

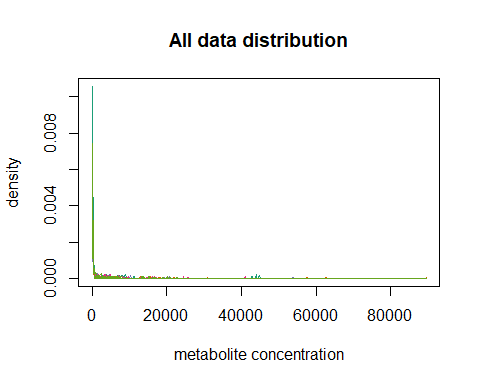
RawData<-read.csv("quantro\_analysis.in.csv")

Let’s begin pre-proccessing by selecting only the metabolite concentraions.

PreData<-t(RawData[,-c(1,2)])

Let’s look at a simple histogram of all the metabolite concentrations. We also transpose the RawData because quantro expects samples to be arrangened in columns with the features in rows. Note the first two columns (sample number and group) are removed.

matdensity(PreData, groupFactor = RawData$Label, xlab = "metabolite concentration", ylab = "density",  
 main = "All data distribution", brewer.n = 8, brewer.name = "Dark2")

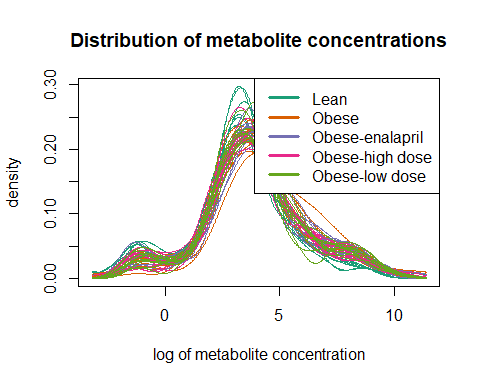


Because metabolite concentrations can vary by orders of magnitute let’s conisder the log of the metabolite concentration. Before we do this we need to remove any zeros from the PreData to avoid inf values. Let’s set any values less than or equal to zero to the square root of the minimum of the feature values across all samples, then take the log.

#replace zeros with sqrt of minimum  
PreData <- as.matrix(t(apply(PreData,1,function(x){  
 x[which(x<=0)]<-sqrt(min(x[which(x>0.)]))  
 return(x)})))  
PreData <- log(PreData)  
#View(PreData)

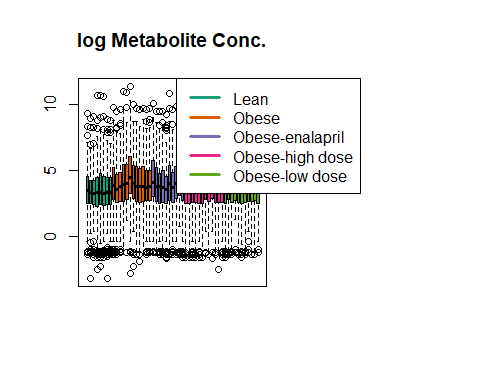
Let’s now plot the density of the data arranged by group to see if we can’t notice any targeted or global differences in the data.

matdensity(PreData, groupFactor = RawData$Label, xlab = "log of metabolite concentration", ylab = "density",  
 main = "Distribution of metabolite concentrations", brewer.n = 8, brewer.name = "Dark2")  
legend(x="topright",legend=unique(RawData$Label),col = c(1,2,3,4,5), lty = 1, lwd =3)



Off hand I’d say there are global differences in the data. One sample from the obese group has particularly high density in the region around log(conc.) ~ 6. Let’s look at the box plot now.

par(mar=c(5.1, 4.1, 4.1, 11.1), xpd=TRUE)  
matboxplot(PreData, groupFactor =RawData$Label, xaxt = "n", main = "log Metabolite Conc.")  
legend(x="topright",inset = c(-0.5,0),legend=unique(RawData$Label),col = c(1,2,3,4,5), lty = 1, lwd =3)



All the data seem to fall within the error bars, however the “Obese” group has some variability in the mean. The “lean” group seems to have on average a lower mean with a tight distribution. Qualitatively, it looks like there is disparity between the variability within groups and the variability between groups. Let’s use quantro to quantify these two types of variance.

## Main Quantro Analysis

qtest <- quantro(object=PreData, groupFactor = RawData$Label, B=500)

## [quantro] Average medians of the distributions are   
## not equal across groups.

## [quantro] Calculating the quantro test statistic.

## [quantro] Starting permutation testing.

## [quantro] Using a single core (backend: doSEQ,   
## version: 1.4.4).

qtest

## quantro: Test for global differences in distributions  
## nGroups: 5   
## nTotSamples: 53   
## nSamplesinGroups: 8 12 10 12 11   
## anovaPval: 0   
## quantroStat: 2.8253   
## quantroPvalPerm: < 0.002

We can confirm what we saw in the boxplot. Anova states average medians are not equal across group, so global shift exists. This can be due to technical or biological sources. The quantroStat value tells us that the vaiability of distributions between the groups is 2.8253 time larger than the variability of the distributions within the groups. Quantile normalization is NOT appropriate unless we can definately say that the source of the global variation is coming from technical, not biological, sources. Look’s like the anova pvalue is 0. This is odd; let’s look at the anova results in detail.

anova(qtest)

## Analysis of Variance Table  
##   
## Response: objectMedians  
## Df Sum Sq Mean Sq F value Pr(>F)   
## groupFactor 4 1.6020 0.40051 13.367 2.151e-07 \*\*\*  
## Residuals 48 1.4381 0.02996   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

OK, pvalue is not zero; it is just really small. Good. Let’ now look at all the other quantro output.

MSbetween(qtest)

## [1] 0.1755335

MSwithin(qtest)

## [1] 0.06212914

Mean suared error (MSE) between groups is 0.1755 while MSE within groups is 0.0621.

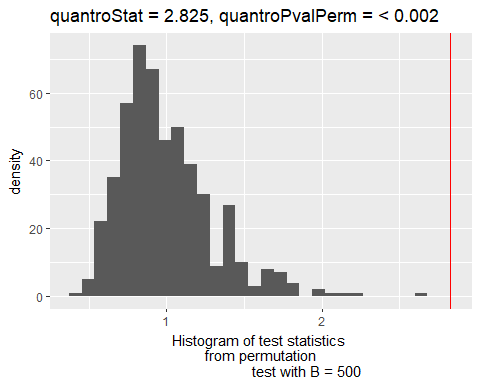
We tested 500 permutations samples in this analysis to quantify the significance of the quantroStat.

quantroPvalPerm(qtest)

## [1] 0

The above value is the p-value associated with the proportion of times the test statistics from the permutation samples were larger than quantroStat. So in this case we are very confident the quantroStat is a good measure of the variability because the pvalue is small 0.002 (see following plot).

quantroPlot(qtest)



A histogram containing the null test statistics when using boostrapped samples. The red line is the observed test statistic quantroStat. Look’s like we are very far in the tail end of the density. It is likely that this p-value can get much smaller if we increase the resolution of the curve by increasing the number of permutation samples. This would only be academic as we are already sufficiently confident in the quantroStat value.

## Conclusion

Global differences in the distribution of the data exists. Quantile normalization will wash out this information and is therefore not appropriate.