NASA quantro analysis - Remove HDT-1 outlier

Daniel Montemayor

Jan 16, 2019

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

## Summary

Global differences in the distribution are not strongly apparent. Quantile normalization is appropriate. Removing an outlier sample from the HDT-1 group may have strong effects on these results, however, I have no reason to remove this sample.

## Setup requirements

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

## Loading required package: rmarkdown

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.

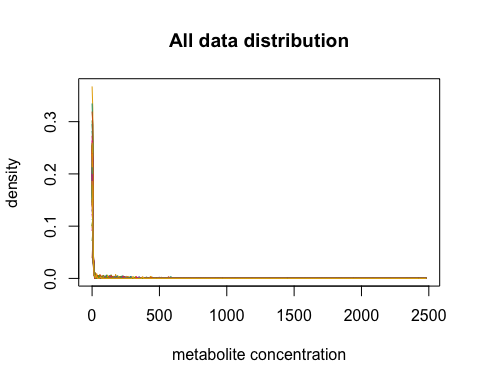
RawData<-read.csv("quantro\_analysis.in.csv")

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$Group, xlab = "metabolite concentration", ylab = "density",  
 main = "All data distribution", brewer.n = 8, brewer.name = "Dark2")

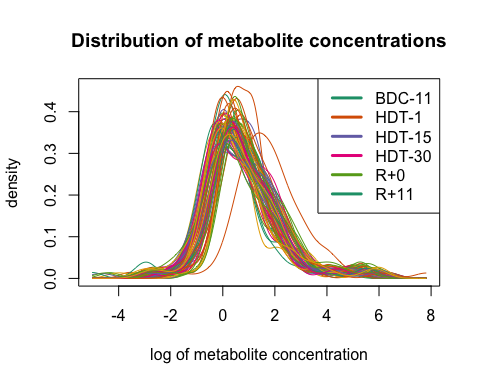


Because metabolite concentrations can vary by orders of magnitute, let’s conisder the log of the metabolite concentration. Before we do this we need to remove any zeros from the PreData to avoid inf values. Let’s set any values less than or equal to zero to the square root of the minimum of the feature 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$Group, xlab = "log of metabolite concentration", ylab = "density",  
 main = "Distribution of metabolite concentrations", brewer.n = 8, brewer.name = "Dark2")  
legend(x="topright",legend=unique(RawData$Group),col = c(1,2,3,4,5), lty = 1, lwd =3)



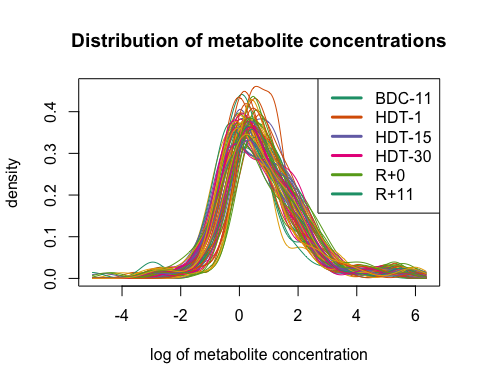
Off hand, I’d say there are no global differences in the data. Although, one sample from the HDT-1 has particularly high density in the region around log(conc.) ~ 2. I see a variability within groups on the same order of magnitude as variability across groups. This suggests quantile normalization would be appropriate.

Let’s continue the analysis with the HDT-1 outlier sample removed.

#remove outlier  
PreData <- PreData[,!RawData$Converted.name=="F2"]  
RawData <- RawData[!RawData$Converted.name=="F2",]

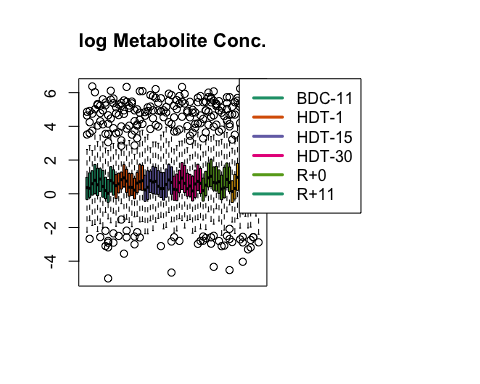
We confirm the outlier is removed by reploting the distributions.

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



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$Group, xaxt = "n", main = "log Metabolite Conc.")  
legend(x="topright",inset = c(-0.5,0),legend=unique(RawData$Group),col = c(1,2,3,4,5), lty = 1, lwd =3)



All the data seem to fall within the same error bars. With the outlier HDT-1 sample removed, it looks like there is no 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$Group, B=5000)

## [quantro] Average medians of the distributions are   
## 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: 6   
## nTotSamples: 66   
## nSamplesinGroups: 11 11 11 11 11 11   
## anovaPval: 0.76835   
## quantroStat: 1.9513   
## quantroPvalPerm: 0.0058

We can confirm what we saw in the boxplot. Anova states average medians are equal across groups, so global shift does not exist. The quantroStat value tells us that the variability of distributions between the groups is only 1.95 times larger than the variability of the distributions within the groups. This suggest some but not strong global changes are apparent. **Quantile normalization is appropriate.** 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 5 0.1576 0.031520 0.509 0.7683  
## Residuals 60 3.7156 0.061927

The Pr(>F) value is the p-value of the calculattion. A value of 0.76835 tell us that we cannot make any confidant claim of any global shift. Good. Let’s now look at all the other quantro output.

MSbetween(qtest)

## [1] 0.07350922

MSwithin(qtest)

## [1] 0.03767189

Mean suared error (MSE) between groups is 0.073 while MSE within groups is 0.038. This is where we get the quantro stat of 1.95.

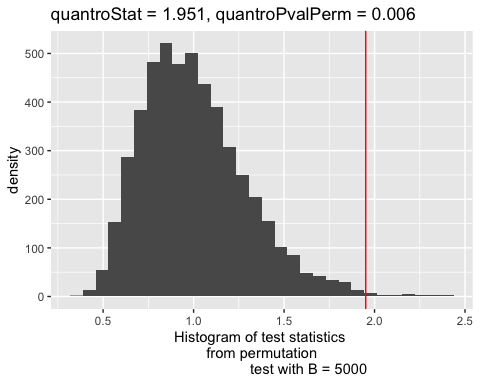
We tested 5000 permutations samples in this analysis to quantify the significance of the quantroStat.

quantroPvalPerm(qtest)

## [1] 0.0058

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 0.008 (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 have significance with p-value = 0.006. Removing the outlier sample from the HDT-1 group has helped the statistics.

## Conclusion

Global differences in the distribution are not strong. Quantile normalization is appropriate.