Differential Expression with R

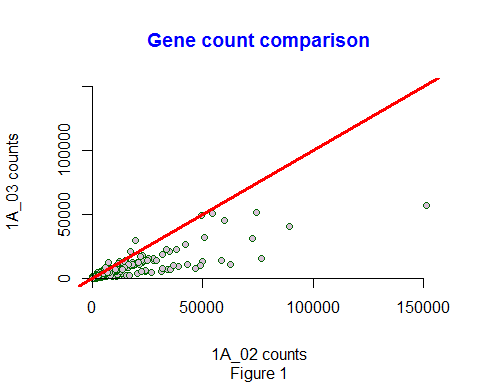
Devin Dinwiddie

January 15, 2018

read\_totals=read.table("A7reads.tsv",header = F) # import read total into data frame  
read\_totals$V1= substring(read\_totals$V1,1,5) # clean up file name remove .fq.gz  
colnames(read\_totals)=c("file name","total reads") # rename the columns  
  
counts=read.delim("Gacu\_gut\_counts.tsv",row.names = 1,stringsAsFactors = F) # read gene counts data in df  
countsum=data.frame(apply(counts,2,sum)) # sum the gene counts data across files ie columns (2)  
colnames(countsum)="sumOfExonReadCounts" # rename the columns for clarity  
  
#determine the percentage of reads that mapped to exons  
countProp=data.frame(read\_totals$`file name`,(countsum$sumOfExonReadCounts/read\_totals$`total reads`)\*100)  
colnames(countProp)=c("file name","percent mapping to exon") # rename the columns for clarity  
head(countProp)

## file name percent mapping to exon  
## 1 1A\_02 55.00006  
## 2 1A\_03 56.51105  
## 3 3A\_01 47.63764  
## 4 3A\_02 63.27541  
## 5 3C\_01 62.94352  
## 6 3C\_04 63.11996

#plot 1A\_02 vs 1A\_03 to compare read lengths  
plot(counts$X1A\_02,counts$X1A\_03,xlab= "1A\_02 counts",ylab="1A\_03 counts",axes=F,col="darkgreen",col.lab="black",main="Gene count comparison",col.main='blue',sub="Figure 1",cex=1,pch=21,bg="thistle",ylim = c(0,150000))  
axis(1,pos=0)  
axis(2,pos=0)  
abline(0,1,col='red',lwd=3)



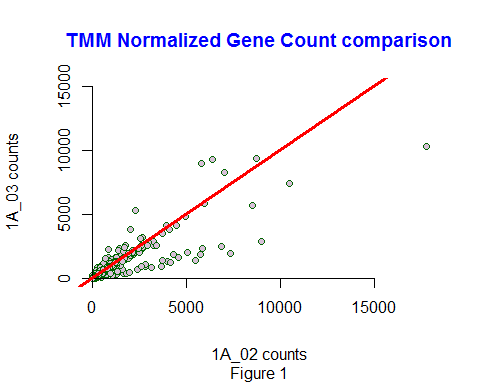
library(edgeR) # load the edge r library

## Warning: package 'edgeR' was built under R version 3.3.2

## Loading required package: limma

## Warning: package 'limma' was built under R version 3.3.3

dge = DGEList(counts=counts) #store raw counts in a dgelist  
dge = calcNormFactors(dge) # calculate tmm normalization factors  
TMMCPM = as.data.frame(cpm(dge,normalized.lib.sizes=TRUE)) #Apply the TMM normalization factors and adjust for library size by converting to copies per million (cpm)  
  
#plot normalized gene count data same samples as previous plot  
plot(TMMCPM$X1A\_02,TMMCPM$X1A\_03,xlab= "1A\_02 counts",ylab="1A\_03 counts",axes=F,col="darkgreen",col.lab="black",main="TMM Normalized Gene Count comparison",col.main='blue',sub="Figure 1",cex=1,pch=21,bg="thistle",ylim = c(0,15000))  
axis(1,pos=0)  
axis(2,pos=0)  
abline(0,1,col='red',lwd=3)



dge.er = DGEList(counts=counts) #new dge list  
keep.dge.er = rowSums(cpm(dge.er)>=1) >= 8 # only keep genes seen at least 2 cpm in at least 8 sampls  
dge.er = dge.er[keep.dge.er,] #update dge list with low level genes factored out  
  
#how many genes are we dealing with now  
nrow(dge.er)

## [1] 16877

dge.er = calcNormFactors(dge.er) #calculate TMM normalizaton factor  
  
Targets.er=read.delim("Gacu\_gut\_metadata.tsv") #read in the meta data  
  
PopTrt = paste(Targets.er$Population,Targets.er$Treatment, sep=".") #combine pop and treatment for contrasts  
PopTrt

## [1] "Bt.CV" "Bt.CV" "Bt.CV" "Bt.CV" "Bt.GF" "Bt.GF" "Bt.GF" "Bt.GF"  
## [9] "Bt.CV" "Bt.CV" "Bt.CV" "Bt.CV" "Bt.GF" "Bt.GF" "Bt.GF" "Bt.GF"  
## [17] "RS.CV" "RS.CV" "RS.CV" "RS.CV" "RS.CV" "RS.CV" "RS.GF" "RS.GF"  
## [25] "RS.GF" "RS.GF" "RS.GF" "RS.GF" "RS.CV" "RS.CV" "RS.GF" "RS.GF"

design.er = model.matrix(~0 + PopTrt) #set up the design error  
design.er

## PopTrtBt.CV PopTrtBt.GF PopTrtRS.CV PopTrtRS.GF  
## 1 1 0 0 0  
## 2 1 0 0 0  
## 3 1 0 0 0  
## 4 1 0 0 0  
## 5 0 1 0 0  
## 6 0 1 0 0  
## 7 0 1 0 0  
## 8 0 1 0 0  
## 9 1 0 0 0  
## 10 1 0 0 0  
## 11 1 0 0 0  
## 12 1 0 0 0  
## 13 0 1 0 0  
## 14 0 1 0 0  
## 15 0 1 0 0  
## 16 0 1 0 0  
## 17 0 0 1 0  
## 18 0 0 1 0  
## 19 0 0 1 0  
## 20 0 0 1 0  
## 21 0 0 1 0  
## 22 0 0 1 0  
## 23 0 0 0 1  
## 24 0 0 0 1  
## 25 0 0 0 1  
## 26 0 0 0 1  
## 27 0 0 0 1  
## 28 0 0 0 1  
## 29 0 0 1 0  
## 30 0 0 1 0  
## 31 0 0 0 1  
## 32 0 0 0 1  
## attr(,"assign")  
## [1] 1 1 1 1  
## attr(,"contrasts")  
## attr(,"contrasts")$PopTrt  
## [1] "contr.treatment"

colnames(design.er)

## [1] "PopTrtBt.CV" "PopTrtBt.GF" "PopTrtRS.CV" "PopTrtRS.GF"

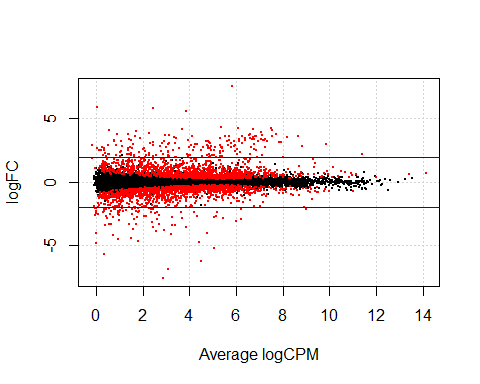
#estimate dispersion parameter for modleing  
dge.er = estimateGLMCommonDisp(dge.er, design.er) #serial estimate common dispersion  
dge.er = estimateGLMTrendedDisp(dge.er, design.er) #estimate trended dispersion  
dge.er = estimateGLMTagwiseDisp(dge.er, design.er) #estimate tagwise dispersion  
fit.er = glmFit(dge.er, design.er)# fit generalized linear model with dispersion estimates  
  
lrtPop = glmLRT(fit.er, contrast=c(-.5,-.5,.5,.5)) #test pop effect contrast the average across sample groups  
topTags(lrtPop) #genes with lowest corrected p values

## Coefficient: -0.5\*PopTrtBt.CV -0.5\*PopTrtBt.GF 0.5\*PopTrtRS.CV 0.5\*PopTrtRS.GF   
## logFC logCPM LR PValue  
## ENSGACG00000020944 7.371826 5.839152 1041.4029 1.798659e-228  
## ENSGACG00000016064 -4.031369 4.764007 580.7140 2.622799e-128  
## ENSGACG00000009551 -2.955664 6.234024 448.0280 1.937669e-99  
## ENSGACG00000019588 2.873130 5.298303 383.9250 1.739524e-85  
## ENSGACG00000018602 2.484802 4.145942 319.0296 2.356500e-71  
## ENSGACG00000020259 2.532350 4.946969 254.0133 3.463700e-57  
## ENSGACG00000007674 -2.823229 5.452999 244.9621 3.256864e-55  
## ENSGACG00000017175 -3.552164 3.874009 241.7399 1.641835e-54  
## ENSGACG00000003825 -2.708820 5.147458 235.8815 3.110017e-53  
## ENSGACG00000013202 -4.710729 4.409334 214.4213 1.490420e-48  
## FDR  
## ENSGACG00000020944 3.035596e-224  
## ENSGACG00000016064 2.213249e-124  
## ENSGACG00000009551 1.090068e-95  
## ENSGACG00000019588 7.339487e-82  
## ENSGACG00000018602 7.954129e-68  
## ENSGACG00000020259 9.742810e-54  
## ENSGACG00000007674 7.852299e-52  
## ENSGACG00000017175 3.463655e-51  
## ENSGACG00000003825 5.831973e-50  
## ENSGACG00000013202 2.515381e-45

write.table(topTags(lrtPop, n=16877), 'edgeR\_PopLRT.txt',sep='\t') #write test results to outfile   
de.er = decideTestsDGE(lrtPop, p=0.05)   
summary(de.er)#summary of genes differentially expressed by pop

## [,1]   
## -1 1320  
## 0 14125  
## 1 1432

detags.er = rownames(fit.er)[as.logical(de.er)]  
#plot in a smearplot  
plotSmear(lrtPop, de.tags=detags.er)  
abline(h = c(-2, 2), col = "blue")



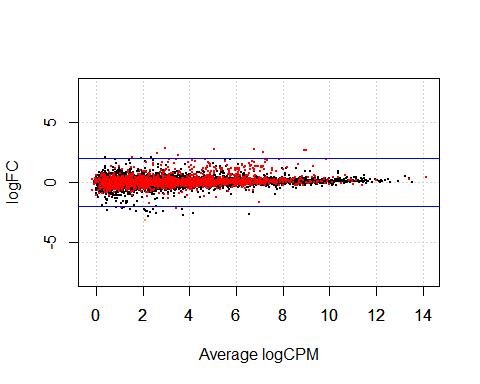
# perform the same calculations for treatment  
#to test for the treatment (microbiota) effects we contrast the #average of Bt.CV and RS.CV the average of Bt.GF and RS.GF  
lrtTrt = glmLRT(fit.er, contrast=c(-.5,.5,-.5,.5))  
topTags(lrtTrt) #look for genes with lowest FDR corrected Pval

## Coefficient: -0.5\*PopTrtBt.CV 0.5\*PopTrtBt.GF -0.5\*PopTrtRS.CV 0.5\*PopTrtRS.GF   
## logFC logCPM LR PValue FDR  
## ENSGACG00000017166 -2.676951 6.5470971 43.48027 4.282564e-11 7.227684e-07  
## ENSGACG00000014415 -1.146976 4.6298272 28.06459 1.173330e-07 9.901146e-04  
## ENSGACG00000012927 -1.239857 5.1617038 26.16113 3.140808e-07 1.766914e-03  
## ENSGACG00000011948 -2.755730 3.7208664 24.37873 7.913781e-07 3.014997e-03  
## ENSGACG00000006658 -1.766239 0.4235724 24.14554 8.932266e-07 3.014997e-03  
## ENSGACG00000013747 -1.266271 4.4723561 23.75666 1.093162e-06 3.074882e-03  
## ENSGACG00000008349 2.731992 6.7633382 22.00049 2.725806e-06 6.250371e-03  
## ENSGACG00000010918 -3.547505 2.0935274 21.72386 3.148505e-06 6.250371e-03  
## ENSGACG00000021894 -2.585360 4.1391874 21.61455 3.333136e-06 6.250371e-03  
## ENSGACG00000015537 -1.077423 3.8982091 20.89959 4.839974e-06 8.136386e-03

write.table(topTags(lrtTrt, n=16877), 'edgeR\_TrtLRT.txt',sep='\t') #write to a table  
de.er = decideTestsDGE(lrtTrt, p=0.05)  
summary(de.er) #sumarize the differential exppression by treat

## [,1]   
## -1 40  
## 0 16828  
## 1 9

#create a smear plot for the treatment comparison  
plotSmear(lrtTrt,de.tags=detags.er,yaxt='n',ylim=c(-8,8))  
axis(1)  
axis(2,at=c(-5,0,5),labels = c(-5,0,5))  
abline(h = c(-2, 2), col = "blue")



The treatment comparison smear plot has the differentially expressed genes (red) aligning around 0, The two plots together indicate a much higher frequency of DE genes amongst the stickleback populations and not a high frequency of DE genes in the different microbiota treatments.

#Testing effects of Population and Treatment on differential expression   
#using the general linear model approach in limma (with voom())  
dge.lv <- DGEList(counts=counts) #new dge list  
  
keep.dge.lv = rowSums(cpm(dge.er)>=1) >= 8 # only keep genes seen at least 2 cpm in at least 8 sampls  
dge.lv = dge.lv[keep.dge.lv,] #update dge list with low level genes factored out  
  
#how many genes are we dealing with now  
nrow(dge.lv)

## [1] 22362

dge.lv = calcNormFactors(dge.lv) #calculate TMM normalizaton factor  
  
Targets.lv=read.delim("Gacu\_gut\_metadata.tsv") #read in the meta data  
  
PopTrt.lv = paste(Targets.lv$Population,Targets.lv$Treatment, sep=".") #combine pop and treatment for contrasts  
PopTrt.lv

## [1] "Bt.CV" "Bt.CV" "Bt.CV" "Bt.CV" "Bt.GF" "Bt.GF" "Bt.GF" "Bt.GF"  
## [9] "Bt.CV" "Bt.CV" "Bt.CV" "Bt.CV" "Bt.GF" "Bt.GF" "Bt.GF" "Bt.GF"  
## [17] "RS.CV" "RS.CV" "RS.CV" "RS.CV" "RS.CV" "RS.CV" "RS.GF" "RS.GF"  
## [25] "RS.GF" "RS.GF" "RS.GF" "RS.GF" "RS.CV" "RS.CV" "RS.GF" "RS.GF"

design.lv = model.matrix(~0 + PopTrt) #set up the design error  
design.lv

## PopTrtBt.CV PopTrtBt.GF PopTrtRS.CV PopTrtRS.GF  
## 1 1 0 0 0  
## 2 1 0 0 0  
## 3 1 0 0 0  
## 4 1 0 0 0  
## 5 0 1 0 0  
## 6 0 1 0 0  
## 7 0 1 0 0  
## 8 0 1 0 0  
## 9 1 0 0 0  
## 10 1 0 0 0  
## 11 1 0 0 0  
## 12 1 0 0 0  
## 13 0 1 0 0  
## 14 0 1 0 0  
## 15 0 1 0 0  
## 16 0 1 0 0  
## 17 0 0 1 0  
## 18 0 0 1 0  
## 19 0 0 1 0  
## 20 0 0 1 0  
## 21 0 0 1 0  
## 22 0 0 1 0  
## 23 0 0 0 1  
## 24 0 0 0 1  
## 25 0 0 0 1  
## 26 0 0 0 1  
## 27 0 0 0 1  
## 28 0 0 0 1  
## 29 0 0 1 0  
## 30 0 0 1 0  
## 31 0 0 0 1  
## 32 0 0 0 1  
## attr(,"assign")  
## [1] 1 1 1 1  
## attr(,"contrasts")  
## attr(,"contrasts")$PopTrt  
## [1] "contr.treatment"

colnames(design.lv) #print column nams

## [1] "PopTrtBt.CV" "PopTrtBt.GF" "PopTrtRS.CV" "PopTrtRS.GF"

#use voom to generate precision weights  
v.lv = voom(dge.lv, design.lv)  
#fit the full general linear model, which includes the voom() precision weights  
fit.lv = lmFit(v.lv, design.lv)  
  
#compute contrasts from linear model fit  
fit.lv.Pop = contrasts.fit(fit.lv,contrast=c(-.5,-.5,.5,.5))  
#Ebayes statistics for Differential Expression   
fit.lv.Pop=eBayes(fit.lv.Pop)  
#look at all the genes with the lowest FDR-corrected p-values  
topTable(fit.lv.Pop)

## logFC AveExpr t P.Value adj.P.Val  
## ENSGACG00000009551 -2.969911 5.547030 -27.56463 2.966766e-23 6.634282e-19  
## ENSGACG00000020944 7.484019 3.047776 25.57878 2.697261e-22 3.015807e-18  
## ENSGACG00000016064 -4.076550 3.584227 -23.55565 3.020668e-21 2.251606e-17  
## ENSGACG00000019588 2.913860 4.620351 22.12201 1.878596e-20 1.050229e-16  
## ENSGACG00000018602 2.464053 3.617968 19.13924 1.207194e-18 4.499211e-15  
## ENSGACG00000015400 3.980240 -1.555844 19.55755 6.518282e-19 2.915236e-15  
## ENSGACG00000007674 -2.866965 4.760158 -16.90140 4.010574e-17 1.281206e-13  
## ENSGACG00000003825 -2.781527 4.498539 -16.79658 4.769341e-17 1.333150e-13  
## ENSGACG00000017515 2.284108 3.434233 16.02935 1.742171e-16 4.328714e-13  
## ENSGACG00000020259 2.502892 4.390563 15.46130 4.694713e-16 1.049832e-12  
## B  
## ENSGACG00000009551 42.68813  
## ENSGACG00000020944 38.04800  
## ENSGACG00000016064 37.39062  
## ENSGACG00000019588 36.48918  
## ENSGACG00000018602 32.34508  
## ENSGACG00000015400 30.35555  
## ENSGACG00000007674 29.05967  
## ENSGACG00000003825 28.87573  
## ENSGACG00000017515 27.54892  
## ENSGACG00000020259 26.64364

#write results to a table  
write.table(topTable(fit.lv.Pop, n=16877),'limma\_PopLRT.txt', sep='\t')  
  
#summarize how many genes are differentially expressed by Populaton   
#using an FDR = 0.05.   
subsetPop <- subset(topTable(fit.lv.Pop, n=16877), topTable(fit.lv.Pop, n=16877)[,5] <= 0.05)

#Treatment Contrast   
fit.lv.Trt = contrasts.fit(fit.lv,contrast=c(-.5,.5,-.5,.5))  
fit.lv.Trt = eBayes(fit.lv.Trt)  
#To look at the genes with the lowest FDR-corrected p-values:  
topTable(fit.lv.Trt)

## logFC AveExpr t P.Value adj.P.Val  
## ENSGACG00000014415 -1.1206052 4.3221428 -4.920131 2.747874e-05 0.2842080  
## ENSGACG00000017166 -1.8859320 5.4034131 -4.672428 5.565241e-05 0.2842080  
## ENSGACG00000017357 -0.6848217 2.8387850 -4.625743 6.354709e-05 0.2842080  
## ENSGACG00000012927 -1.1285196 4.7955226 -4.292379 1.630604e-04 0.3979438  
## ENSGACG00000011673 -0.7592235 5.4745335 -4.265250 1.759747e-04 0.3979438  
## ENSGACG00000010022 -0.7004164 1.9092263 -4.640251 6.098159e-05 0.2842080  
## ENSGACG00000010159 -0.7114827 2.2645527 -4.427987 1.112712e-04 0.3979438  
## ENSGACG00000003073 0.5138266 4.6336934 4.172737 2.280614e-04 0.3979438  
## ENSGACG00000006466 -0.4809835 2.4532569 -4.340722 1.423231e-04 0.3979438  
## ENSGACG00000006658 -1.8754930 -0.4635065 -5.309521 9.032066e-06 0.2019750  
## B  
## ENSGACG00000014415 2.2970641  
## ENSGACG00000017166 1.7778138  
## ENSGACG00000017357 1.2251561  
## ENSGACG00000012927 0.8358659  
## ENSGACG00000011673 0.7989923  
## ENSGACG00000010022 0.7581093  
## ENSGACG00000010159 0.5472890  
## ENSGACG00000003073 0.5442765  
## ENSGACG00000006466 0.4643127  
## ENSGACG00000006658 0.4500985

#And to write all of the test results to file  
write.table(topTable(fit.lv.Trt, n=16877),'limma\_TrtLRT.txt', sep='\t')  
  
#Summarize how many genes are differentiall expressed by treatment,  
#using an FDR = 0.05 and subset and parse topTable.   
#subsetTrt <- subset(topTable(fit.lv.Trt, n=16877),topTable(fit.lv.Trt, n=16877)[,5] <= 0.05)  
subsetTrt <- subset(topTable(fit.lv.Trt, n=16877), subset = adj.P.Val <= 0.05)

#Compare the overlap in differentially expressed genes between   
#edgeR and limma-voom  
  
# #create a variable to hold the topTags objects.  
topLrtPop<- topTags(lrtPop, n=16877)  
topLrtTrt <- topTags(lrtTrt, n=16877)  
#Create a subset of the topTags for those genes with FDR <= 0.05  
subsetLRTpop <- subset(topLrtPop$table, topLrtPop$table$FDR <= 0.05)   
  
#check that the subset worked   
head(subsetLRTpop)

## logFC logCPM LR PValue  
## ENSGACG00000020944 7.371826 5.839152 1041.4029 1.798659e-228  
## ENSGACG00000016064 -4.031369 4.764007 580.7140 2.622799e-128  
## ENSGACG00000009551 -2.955664 6.234024 448.0280 1.937669e-99  
## ENSGACG00000019588 2.873130 5.298303 383.9250 1.739524e-85  
## ENSGACG00000018602 2.484802 4.145942 319.0296 2.356500e-71  
## ENSGACG00000020259 2.532350 4.946969 254.0133 3.463700e-57  
## FDR  
## ENSGACG00000020944 3.035596e-224  
## ENSGACG00000016064 2.213249e-124  
## ENSGACG00000009551 1.090068e-95  
## ENSGACG00000019588 7.339487e-82  
## ENSGACG00000018602 7.954129e-68  
## ENSGACG00000020259 9.742810e-54

max(subsetLRTpop$FDR)

## [1] 0.0499339

#subset the topTag object to those genes with FDR <= 0.05  
subsetLrtTrt = subset(topLrtTrt$table, subset = topLrtTrt$table$FDR <= 0.05)  
  
#Load the library   
library(VennDiagram)

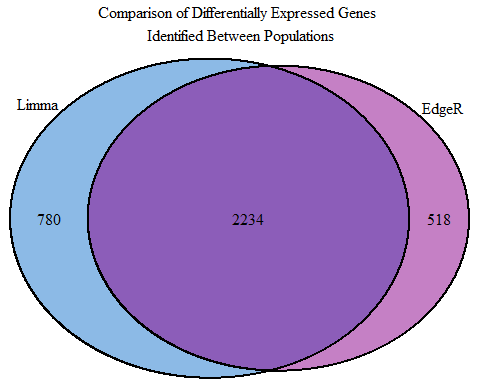
## Warning: package 'VennDiagram' was built under R version 3.3.3

## Loading required package: grid

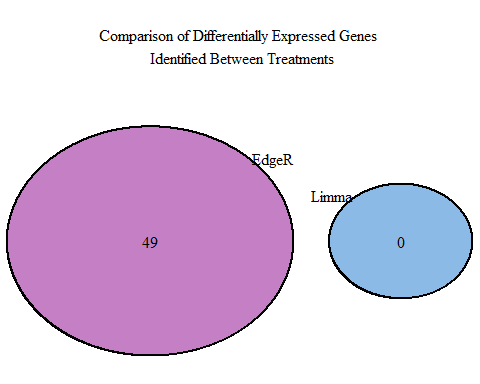
## Loading required package: futile.logger

## Warning: package 'futile.logger' was built under R version 3.3.1

#Make a Venn Diagram to show overlap between EdgeR and Limma approaches.  
#Show overlap of significant genes (FDR = 0.05)  
  
#Variables to use Limma:  
#subsetPop, subsetTrt   
namesLVpop = rownames(subsetPop)  
namesLVtrt = rownames(subsetTrt)   
  
#Variable to use EdgeR:  
#subsetLRTpop, subsetLrtTrt  
namesLrtPop = rownames(subsetLRTpop)  
namesLrtTrt = rownames(subsetLrtTrt)  
  
  
#Characterize the population test.  
#find the set of signif Diff Expressed genes that overlaps between edgeR and limmaVoom approaches  
vennPop = venn.diagram(  
 list(Limma = namesLVpop, EdgeR = namesLrtPop),   
 filename = NULL, fill = c("dodgerblue3", "magenta4"),   
 main = "Comparison of Differentially Expressed Genes \n Identified Between Populations" )  
  
# to get venn di to paste in RMD knit filename=Null above and code below  
grid.newpage()  
grid.draw(vennPop)



#Characterize the treatment test.   
#find the set of significant DE genes between EdgeR and Limma  
  
vennTrt = venn.diagram(list(Limma = namesLVtrt, EdgeR = namesLrtTrt),  
 filename = NULL, fill = c("dodgerblue3", "magenta4"),  
 main = "Comparison of Differentially Expressed Genes \n Identified Between Treatments")  
  
grid.newpage()  
grid.draw(vennTrt)



In the limma-voom approach for treatment comparison 0 DE genes were found while the Edge approach found few (49) no overlap between the two approches for treatment comparison. population comparison found a lot more overlap between the two approches both approaches seem to have found unique DE genes not identified in by other method but the majority of DE genes identified are in the overlap between the two approaches. a lot more DE genes were identifies between populations compared to treatment.