setwd("~/Teaching/SISG/2017/Eight adults/Norm Tutorial")

source("https://bioconductor.org/biocLite.R")

biocLite("edgeR")

library(edgeR)

EIGHT <- read.csv("EIGHTadults\_raw.csv", header=T, row.names="GENE", sep=',')

targets <- read.csv("EIGHTadults\_design.csv", header=T, row.names="Sample", sep=',')

View(EIGHT)

hist(log2(EIGHT$AFO1\_1),main="Distribution of AFO1\_1",xlab="Expression bin")

View(targets)

group <- factor(paste(targets$Ethn,targets$Weight,sep="."))

cbind(targets,Group=group)

y <- DGEList(counts=EIGHT, group=group)

design <- model.matrix(~0+group, data=y$samples)

colnames(design) <- levels(y$samples$group)

head(design)

keep <- rowSums(cpm(y)>2) >= 16

y <- y[keep, , keep.lib.sizes=FALSE]

summary(y$samples)

y <- calcNormFactors(y)

head(y$samples)

plotMDS(y)

y <- estimateGLMCommonDisp(y, design)

y <- estimateGLMTrendedDisp(y, design)

y <- estimateGLMTagwiseDisp(y, design)

plotBCV(y)

boxplot(log2(y$counts), main='Eight Adults', ylab='Gene Expression')

fit <- glmFit(y, design)

lrtW <- glmLRT(fit, contrast=c(-0.5,0.5,-0.5,0.5))

detags <- rownames(topTags(lrtW, p=0.01, adjust="BH", n=20))

plotSmear(lrtW, de.tags=detags, cex=0.8)

summary(decideTestsDGE(lrtW, p=0.05, adjust="BH"))

summary(decideTestsDGE(lrtW, p=0.05, adjust="BH", lfc=0.5))

write.table(lrtW$table, file="de\_weight.csv", sep=",", col.names=NA)

pval <- lrtW$table$PValue

library(qvalue)

qobj <- qvalue(p=pval)

summary(qobj)

hist(qobj$pvalues)

plot(qobj)

write(qobj$qvalues, file="weight\_q.txt", ncolumns=1)

> lrte <- glmLRT(fit, contrast=c(0.5,0.5,-0.5,-0.5))

> detags2 <- rownames(topTags(lrte, p=0.01, adjust="BH", n=20))

> plotSmear(lrte, de.tags=detags2, cex=0.8)

> write.table(lrte$table, file="de\_ethn.csv", sep=",", col.names=NA)

logcpm <- cpm(y, prior.count=2, log=TRUE)

boxplot(logcpm, main='Eight Adults LogCPM', ylab='Gene Expression')

write.table(logcpm, file="logcpm.csv", sep=",", col.names=NA)

(write GENE into first column header)

logEIGHT <- read.csv("logcpm.csv", header=T, row.names="GENE", sep=',')

hist(logEIGHT$AFO1\_1,main="Distribution of AFO1\_1",xlab="Expression bin")

write.table(y$counts, file="reduced.csv", sep=",", col.names=NA)

(write GENE into first column header)

redEIGHT <- read.csv("reduced.csv", header=T, row.names="GENE", sep=',')

plot(log2(redEIGHT$AFO1\_1),logEIGHT$AFO1\_1)

plot(logEIGHT$AFO1\_1,logEIGHT$EFL1\_4)

library("Biobase")

library(pvca)

install.packages("tidyverse") << if “tibble” unavailable

logEIGHT <- as.matrix(read.table("logcpm.csv", header=TRUE, sep=",",row.names =1, as.is=TRUE))

dim(logEIGHT)

all(rownames(targets)==colnames(logEIGHT))

sapply(targets,class)

targets[c(15,20), c("Ethn", "Weight", "Visit")]

phenoData = new("AnnotatedDataFrame", data=targets)

phenoData

EIGHTset <- ExpressionSet(assayData=logEIGHT, phenoData=phenoData)

pct\_threshold = 0.75

batch.factors <- c("Ethn", "Weight")

pvcaObj <- pvcaBatchAssess(EIGHTset, batch.factors, pct\_threshold)

bp <- barplot(pvcaObj$dat, ylab = "Wt average prop var", ylim= c(0,1.1), col = c("blue"), las=2, main="PVCA Estimation")

axis(1, at = bp, labels = pvcaObj$label, cex.axis = 1, las=2)

values = 100\*round(pvcaObj$dat, 2)

text(bp,pvcaObj$dat,labels = values, pos=3, cex = 0.8)

source("https://bioconductor.org/biocLite.R")

biocLite("sva")

library(sva)

phEIGHT <- read.table("EIGHTsva\_design.csv", header=T, row.names="Name", sep=',')

logEIGHT <- read.table("logcpm2.csv", header=TRUE, sep=",",row.names =1, as.is=TRUE)

mod = model.matrix(~as.factor(phEIGHT$weight), data=logEIGHT)

logEIGHT <- as.matrix(logEIGHT)

mod0 = model.matrix(~1,data=phEIGHT)

n.sv = num.sv(logEIGHT,mod,method="leek")

n.sv

svobj = sva(logEIGHT,mod,mod0,n.sv=n.sv)

write.table(svobj$sv, file = "EIGHT\_sv.csv", sep=",", col.names=NA)

pValues = f.pvalue(logEIGHT,mod,mod0)

qValues = p.adjust(pValues,method="BH")

modSv = cbind(mod,svobj$sv)

mod0Sv = cbind(mod0,svobj$sv)

pValuesSv = f.pvalue(logEIGHT,modSv,mod0Sv)

qValuesSv = p.adjust(pValuesSv,method="BH")

write.table(qValuesSv, file = "eight\_adjqv.csv", sep=",", col.names=NA)

plot(qValues,qValuesSv)

batch = phEIGHT$ethn

modcombat = model.matrix(~1, data=phEIGHT)

combat\_data = ComBat(dat=logEIGHT, batch=batch, mod=modcombat, par.prior=TRUE)

pValuesComBat = f.pvalue(combat\_data,mod,mod0)

qValuesComBat = p.adjust(pValuesComBat,method="BH")

write.table(qValuesComBat, file = "eight\_ethnadjqv.csv", sep=",", col.names=NA)

write.table(combat\_data, file = "eight\_ethnadj.csv", sep=",", col.names=NA)

plot(qValuesComBat,qValuesSv)

plot(qValuesComBat,qValues)

phEIGHT <- read.table("EIGHTsva\_design.csv", header=T, row.names="Name", sep=',')

fit1 <- aov(SV1 ~ ethn + weight + visit + person, data=phEIGHT)

summary(fit1)

fit2 <- aov(SV2 ~ ethn + weight + visit + person, data=phEIGHT)

summary(fit2)

fit3 <- aov(SV3 ~ ethn + weight + visit + person, data=phEIGHT)

summary(fit3)

library(snm)

EIGHT.bio = read.csv("EIGHTadults\_bio.csv", header=T, row.names=1)

EIGHT.adj = read.csv("EIGHTadults\_adj.csv", header=T, row.names=1)

EIGHT.int = read.csv("EIGHTadults\_int.csv", header=T)

int.var = EIGHT.int

int.var$Array = as.factor(int.var$Array)

adj.var = model.matrix(~.,EIGHT.adj)

bio.var = model.matrix(~.,EIGHT.bio)

raw.data = as.matrix(logEIGHT)

snm.EIGHT = snm(raw.data,bio.var,adj.var,int.var,rm.adj=FALSE,num.iter=5)

summary(snm.EIGHT)

write.table(snm.EIGHT$norm.dat, file = "EIGHT\_snm.csv", sep=",", col.names=NA)

(write GENE into first column header)

snmEIGHT <- read.csv("EIGHT\_SNM.csv", header=T, row.names="GENE", sep=',')

cpmEIGHT <- read.csv("logcpm.csv", header=T, row.names="GENE", sep=',')

plot(snmEIGHT$AFO1\_1,cpmEIGHT$AFO1\_1)