

# Covariate and surrogate variable-based correction of GTEx data

Warren Anderson  
Joon Yuhl Soh  
Mete Civelek

April 3, 2020

This guide provides code and documentation of analyses from Anderson et al., 2020, *Sex differences in human adipose tissue gene expression and genetic regulation involve adipogenesis*

## Contents

1	Overview	2
2	Analysis of covariate effects in GTEx data	2
3	Covariate correction using surrogate variable analysis	5
4	References	10

## List of Figures

1	Principal component scores for RNA integrity (RIN), age, death circumstances (Hardy Scale) and sex. . . . .	5
2	Principal component scores for collection site and batch type. . . . .	6
3	Principal component scores for covariate-corrected data. . . . .	8
4	Principal component scores for covariate-corrected data. . . . .	9

# 1 Overview

## 2 Analysis of covariate effects in GTEx data

We continue our analysis of the GTEx subcutaneous adipose tissues expression data set using data that were inverse normal transformed on a gene by gene basis. These data are stored in the file named *subq\_gtex\_invNorm.txt* (see vignette *Fig1.GTEx.pdf*). We will use sample covariate data in the *GTEx\_v7\_Annotations\_SampleAttributesDS.txt* file and the subject covariate data in the *GTEx\_v7\_Annotations\_SubjectPhenotypesDS.txt* file downloaded from the GTEx portal as described in the vignette *Fig1.GTEx.pdf*. We also need to download covariate data used in the GTEx eQTL analysis.

```
# website address
https://www.gtexportal.org/home/datasets

# data download, GTEx Analysis V8, Single-Tissue cis-eQTL Data
GTEx_Analysis_v8_eQTL_covariates.tar.gz
gunzip GTEx_Analysis_v8_eQTL_covariates.tar.gz
tar -xvf GTEx_Analysis_v8_eQTL_covariates.tar

# file of interest
Adipose_Subcutaneous.v8.covariates.txt
```

We load in the data of interest and combine all sample and subject covariates.

```
library(dplyr)
library(ggplot2)
library(gridExtra)

# load in data
fname = "GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt"
covar0 = read.table(fname,header=T,sep="\t",stringsAsFactors=F,quote="")
fname = "GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt"
subj0 = read.table(fname,header=T,sep="\t",stringsAsFactors=F,check.names=F)
fname = "subq_gtex_invNorm.txt"
expr0 = read.table(fname,header=T,sep="\t",stringsAsFactors=F,check.names=F)
fname = "Adipose_Subcutaneous.v8.covariates.txt"
ecov0 = read.table(fname,header=T,sep="\t",stringsAsFactors=F,check.names=F) %>% t
names = ecov0[1,]
ecov0 = ecov0[-1,] %>% as.data.frame
names(ecov0) = names

# isolate subcutaneous data in covariate file
covar1 = covar0[covar0$SAMPID %in% names(expr0),]
all(covar1$SAMPID == names(expr0))

# isolate subcutaneous data in subject attribute file
subq_subs = sapply(names(expr0),function(n){
  x = strsplit(n,"-")[[1]]
  return(paste0(x[1],"-",x[2]))
})
subj1 = subj0[subj0$SUBJID %in% subq_subs,]
subj1 = subj1[match(subq_subs,subj1$SUBJID),]
all(names(expr0) == names(subq_subs))
all(subj1$SUBJID == subq_subs)

# verify data organization in eQTL covariate file
all(rownames(ecov0) == subj1$SUBJID)
all(rownames(ecov0) == subq_subs)

# combine all covariates
covar = cbind(covar1,subj1)
```

```
# recode the sex indicator
covar$GENDER[covar$SEX == 1] = "Male"
covar$GENDER[covar$SEX == 2] = "Female"
```

Next, we perform principal components analysis for visualizing systematic influences of particular covariates on the dispersion of the subcutaneous expression data.

```
# implement PCA based on SVD
PC <- prcomp(t(expr0),scale.=T,center=T)
pc_scores = PC$x
pc_loadings = PC$rotation
pc_eigvals = PC$sdev^2
pc_percent = 100 * pc_eigvals / sum(pc_eigvals)
```

Next we plot the principal component projections with annotations based on RNA integrity (RIN), death circumstance, age, sex, collection site, and batch type. Such visualizations provide a basis for comparison to subsequently corrected data (see SVA analysis below). For each subset of the data, indicated by different colors, we plotted the 75% confidence interval enclosing the mean assuming a bivariate normal distribution. The 75% CI was chosen to highlight differences in the PC projections corresponding to each covariate of interest. For the RNA integrity number, we considered samples in the highest and lowest quartiles. Similarly, we show data for only age ranges 30-39 and 60-69 to highlight the influence of age.

```
# plot params
mr = 0.1

# function for generating annotated PCA plots
pca.plot.fun = function(data=NULL,col=NULL,legend.dir=NULL) {
  plt.out = ggplot(data,aes_string(x="PC1",y="PC2",colour=col)) +
    scale_colour_manual(values = c("red","blue","green","cyan","magenta")) +
    geom_point(size=3,alpha=0.7) +
    theme(legend.position="top",legend.direction=legend.dir) +
    theme(plot.margin = unit( c(mr,mr,mr,mr) , "in" ) ) +
    xlab(paste0("PC1 (",round(pc_percent[1],1),"%)" )) +
    ylab(paste0("PC2 (",round(pc_percent[2],1),"%)" )) +
    stat_ellipse(type="norm",level=0.75,size=2)
  return(plt.out)
}

# analyze RIN effects by plotting PC projection annotated according to RIN quartile
quantile_RIN = quantile(covar$SMRIN)
rin = covar$SMRIN
rin[covar$SMRIN==min(covar$SMRIN)] = 1
for(ii in 2:length(quantile_RIN)) {
  ind = which(rin > quantile_RIN[ii-1] & rin <= quantile_RIN[ii])
  rin[ind] = ii-1
}
PCi = data.frame(pc_scores,RIN_quartile=as.factor(rin))
rin.plt = pca.plot.fun(data=subset(PCi,RIN_quartile==c(1,4)),
  col="RIN_quartile",legend.dir="vertical")

# analyze effects of the type of nucleic acid isolation batch
SMNABTCHT = covar$SMNABTCHT
PCi = data.frame(pc_scores,Batch_type=SMNABTCHT)
btch.plt = pca.plot.fun(data=PCi,col="Batch_type",legend.dir="vertical")

# analyze the effects of death circumstances
PCi = data.frame(pc_scores,Hardy_scale=as.factor(covar$DTHHRDY))
```

```

hrdy.plt = pca.plot.fun(data=PCi,col="Hardy_scale",legend.dir="horizontal")

# analyze the effects of the collection site
PCi = data.frame(pc_scores,Collection_site=as.factor(covar$SMCENTER))
coll.plt = pca.plot.fun(data=PCi,col="Collection_site",legend.dir="horizontal")

# annalyze the effects of age
PCi = data.frame(pc_scores,Age=as.factor(covar$AGE))
age.plt = pca.plot.fun(data=subset(PCi,Age==c("30-39","60-69")),
  col="Age",legend.dir="vertical")

# analyze the effects of sex
PCi = data.frame(pc_scores,Sex=as.factor(covar$GENDER))
sex.plt = pca.plot.fun(data=PCi,col="Sex",legend.dir="vertical")

# check sex based on death circumstances
hdyNA <- table(covar$GENDER[is.na(covar$DTHHRDY)]) # Female: 3 Male: 9
hdy0 <- table(covar$GENDER[covar$DTHHRDY == 0]) # Female: 118 Male: 207
hdy1 <- table(covar$GENDER[covar$DTHHRDY == 1]) # Female: 8 Male: 11
hdy2 <- table(covar$GENDER[covar$DTHHRDY == 2]) # Female: 32 Male: 113
hdy3 <- table(covar$GENDER[covar$DTHHRDY == 3]) # Female: 10 Male: 18
hdy4 <- table(covar$GENDER[covar$DTHHRDY == 4]) # Female 23 Male: 29

```

We then ouput the graphs into pdf files.

```

# plot the covariate effects data for RIN etc.
plts = list(rin.plt,hrdy.plt,age.plt,sex.plt)
pdf("PCA_covarRin_uncorrected.pdf", onefile = FALSE)
marrangeGrob(grobs=plts, nrow=2, ncol=2, top=NULL)
dev.off()

# plot the covariate effects data for collect site, etc.
plts = list(coll.plt,btch.plt)
pdf("PCA_covarCol_uncorrected.pdf", onefile = FALSE, height=9, width=4.5)
marrangeGrob(grobs=plts, nrow=2, ncol=1, top=NULL)
dev.off()

```

The results are displayed in Figures 1 and 2. RNA integrity, Age, death circumstances (Hardy Scale) and sex influenced the projection onto the first PC. Note that the Hardy scale value of zero (red) corresponded to subjects that died while on a ventilator. Batch type and collection site showed influences on the second PC as well as the first. Overall, even though the first two PCs accounted for only ~15% of the total variance, the covariates of interest imparted a noticeable influence on the structure of the data.

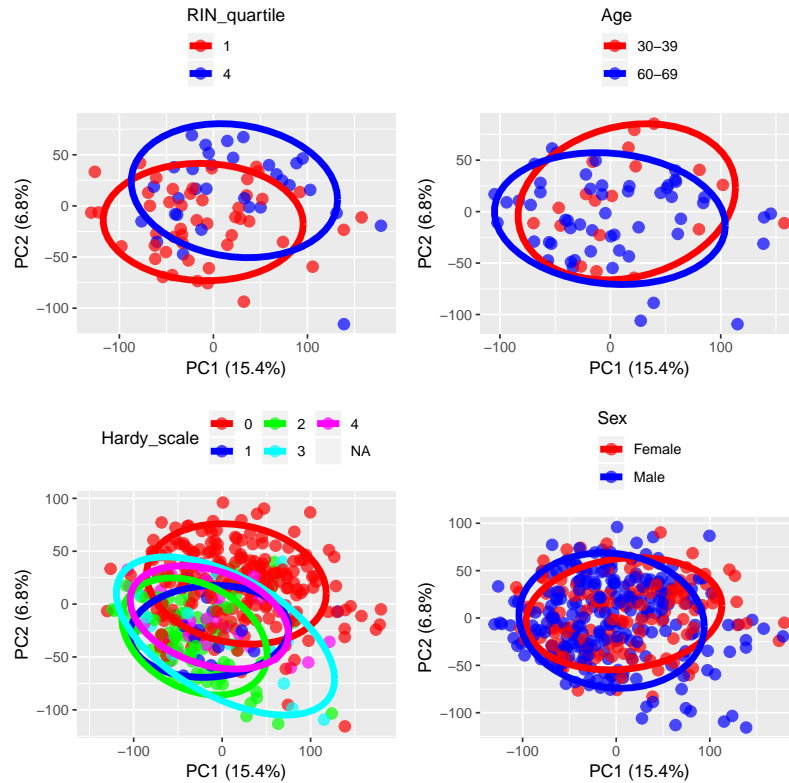


Figure 1: Principal component scores for RNA integrity (RIN), age, death circumstances (Hardy Scale) and sex.

### 3 Covariate correction using surrogate variable analysis

Given the established influence of several known covariates on the genome-wide subcutaneous adipose tissue expression patterns, we wanted to systematically correct the data for both known and unknown covariates, excluding sex - the variable of interest in our analyses. First, we set each age range to a real number indicating the center of the respective range.

```
# age range center
ageCenter = function(vec=NULL){
  age_ranges = unique(vec)
  age_table = matrix(c(0),length(age_ranges),2) %>% as.data.frame
  names(age_table) = c("range","mean")
  for(ii in 1:nrow(age_table)){
    age_mean = strsplit(age_ranges[ii],"-") %>% unlist %>% as.numeric %>% mean
    age_table[ii,1] = age_ranges[ii]
    age_table[ii,2] = age_mean
  }
  ages = age_table$mean[sapply(vec,function(x)which(age_table$range==x))%>%unlist]
  return(ages)
}
covar$AGE = ageCenter(covar$AGE)
```

Now we can implement the surrogate variable analysis (SVA). We set a model including sex, age, RIN, and platform (from the eQTL analysis covariate file) and we infer surrogate variables that should be independent of these known covariates.

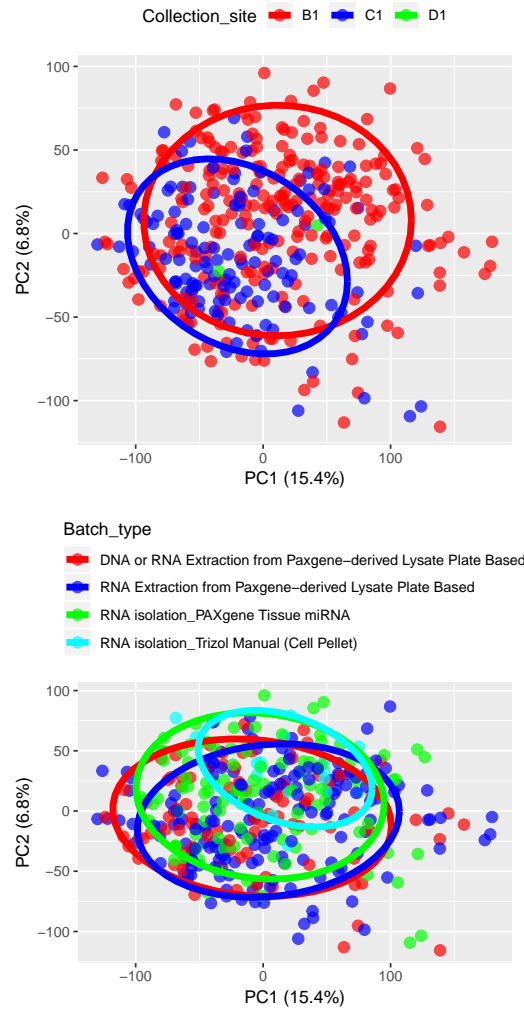


Figure 2: Principal component scores for collection site and batch type.

```
library(sva)

covar$Platform = ecov0$platform

# specify full and null model matrices
modFull = model.matrix(~as.factor(GENDER)+AGE+SMRIN+as.factor(Platform), data=covar)
modNull = model.matrix(~1,data=covar)

# identify the surrogate variables
svadat0 = expr0 %>% data.matrix
LVs = sva(svadat0, modFull, modNull)
LV = LVs$sv
colnames(LV) = paste0("sv",c(1:ncol(LV)))
```

The analysis generated 36 latent variables. Next we correct the expression data for these latent variables along with age, RIN, and platform. We apply multivariate linear and retain the residuals for downstream analysis.

```

# implement regression and keep residuals
# correct for everything other than sex
model0 = covar %>% select(AGE,SMRIN,Platform)
model0$Platform = as.factor(model0$Platform)
model_resid = cbind(model0,LV)

# save the file for future comparison
modelN = covar %>% select(AGE,SMRIN,Platform,GENDER)
modelN$Platform = as.factor(modelN$Platform)
modelN$GENDER[modelN$GENDER == "Male"] = 0
modelN$GENDER[modelN$GENDER == "Female"] = 1
modelN$GENDER = as.factor(modelN$GENDER)
modelS = cbind(LV,modelN)

modelNO <- cbind(colnames(modelN), t(modelN))
colnames(modelNO) <- c("id",covar$SUBJID)
modelSO <- cbind(colnames(modelS), t(modelS))
colnames(modelSO) <- c("id",covar$SUBJID)

fname = "without_sva.txt"
write.table(modelNO,fname,col.names=T,row.names=F,sep="\t",quote=F)
fname = "with_sva.txt"
write.table(modelSO,fname,col.names=T,row.names=F,sep="\t",quote=F)

fitdat = expr0 %>% t
fit = apply(fitdat,2,function(x){
  dat = as.data.frame(cbind(x,model_resid))
  fit_regdat = lm(x~.,data=dat)
  return(fit_regdat$residuals)
})
resids_expr = t(fit)

```

Given the corrected data, we can re-evaluate the influences of known covariates by applying PCA as above.

```

# implement PCA based on SVD
PC <- prcomp(t(resids_expr),scale.=T,center=T)
pc_scores = PC$x
pc_loadings = PC$rotation
pc_eigvals = PC$sdev^2
pc_percent = 100 * pc_eigvals / sum(pc_eigvals)

# analyze RIN effects by plotting PC projection annotated according to RIN quartile
PCi = data.frame(pc_scores,RIN_quartile=as.factor(rin))
rin.plt = pca.plot.fun(data=subset(PCi,RIN_quartile==c(1,4)),
  col="RIN_quartile",legend.dir="vertical")

# analyze effects of the type of nucleic acid isolation batch
SMNABTCHT = covar$SMNABTCHT
PCi = data.frame(pc_scores,Batch_type=SMNABTCHT)
btch.plt = pca.plot.fun(data=PCi,col="Batch_type",legend.dir="vertical")

# analyze the effects of death circumstances
PCi = data.frame(pc_scores,Hardy_scale=as.factor(covar$DTHHRDY))
hrdy.plt = pca.plot.fun(data=PCi,col="Hardy_scale",legend.dir="horizontal")

# analyze the effects of the collection site
PCi = data.frame(pc_scores,Collection_site=as.factor(covar$SMCENTER))

```

```

coll.plt = pca.plot.fun(data=PCi,col="Collection_site",legend.dir="horizontal")

# annalyze the effects of age
PCi = data.frame(pc_scores,Age=as.factor(subj1$AGE))
age.plt = pca.plot.fun(data=subset(PCi,Age==c("30-39","60-69")),
col="Age",legend.dir="vertical")

# analyze the effects of sex
PCi = data.frame(pc_scores,Sex=as.factor(covar$GENDER))
sex.plt = pca.plot.fun(data=PCi,col="Sex",legend.dir="vertical")

# plot the covariate effects data for RIN etc
plts = list(rin.plt,hrdy.plt,age.plt,sex.plt)
pdf("PCA_covarRin_corrected.pdf", onefile = FALSE)
marrangeGrob(grobs=plts, nrow=2, ncol=2, top=NULL)
dev.off()

# plot the covariate effects data for RIN etc
plts = list(coll.plt,btch.plt)
pdf("PCA_covarCol_corrected.pdf", onefile = FALSE, height=9, width=4.5)
marrangeGrob(grobs=plts, nrow=2, ncol=1, top=NULL)
dev.off()

```

The results of applying PCA to the correct data are shown in Figures 3 and 4. The variance captured was 1.39% and 1.21% for PC1 and PC2, respectively. Covariates that previously contributed noticeably to the projection were no longer influential. However, the influence of sex was accentuated. Finally, we output the corrected data and organized covariate matrix for downstream analysis.

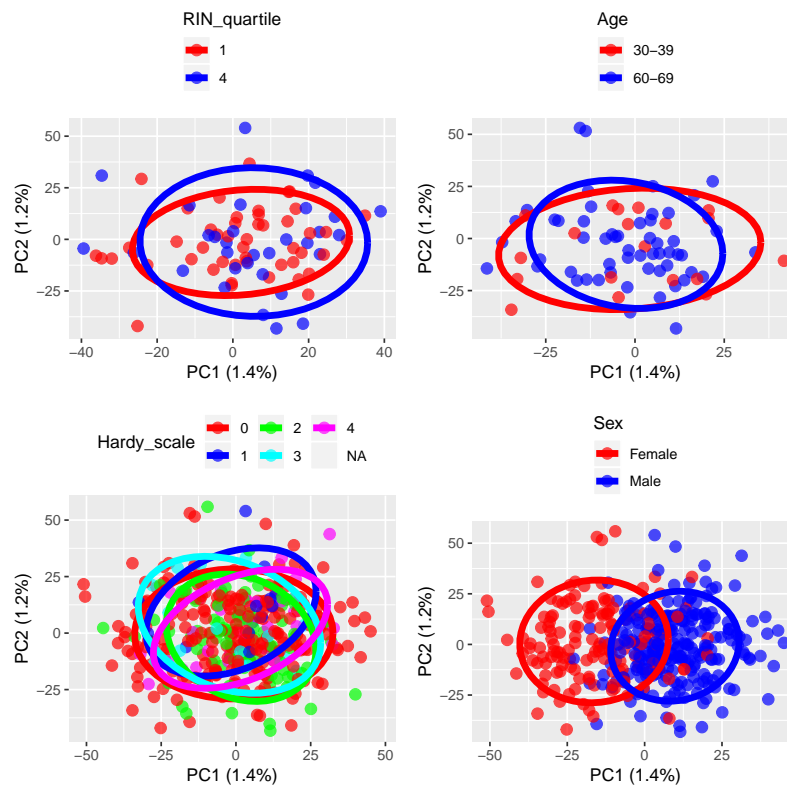


Figure 3: Principal component scores for covariate-corrected data.

Now we export the final data.



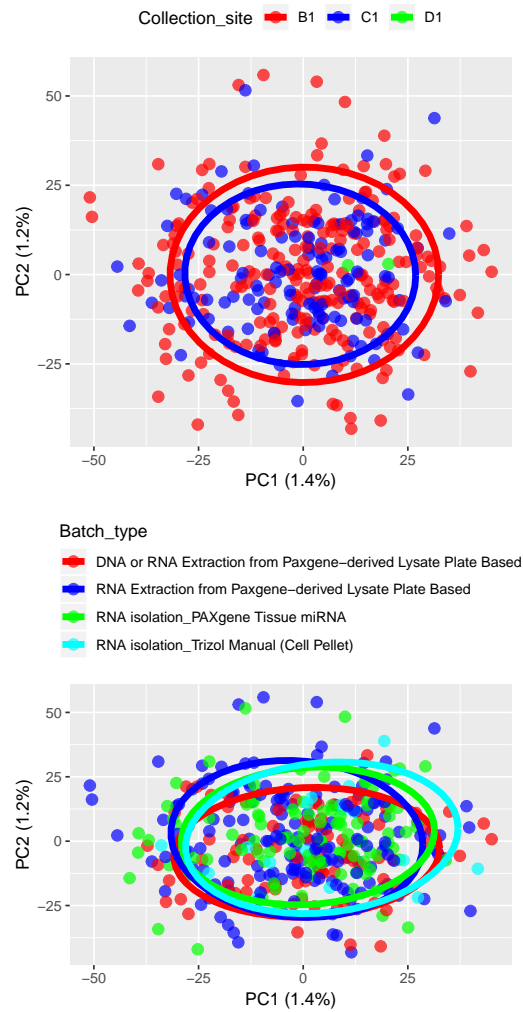


Figure 4: Principal component scores for covariate-corrected data.

```
# output data
fname = "gtex_subq_expr_sva.txt"
write.table(resids_expr,fname,col.names=T,row.names=T,sep="\t",quote=F)

# save the file without a problematic column that we are not interested in
covar <- covar[,c(seq(1,3),seq(5,dim(covar)[2]))]
fname = "gtex_subq_covars.txt"
write.table(covar,fname,col.names=T,row.names=F,sep="\t",quote=F)
```