PCA analysis

Rachel Richardson

November 18, 2018

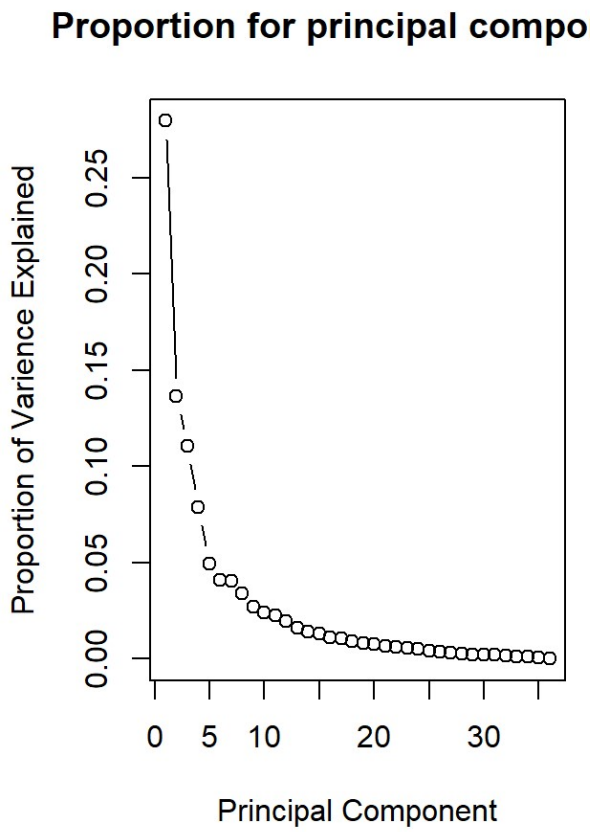
In order to determine if the lipid profiles are different between experimental groups, we've run a PCA across lipid data. (Also to check for batch effects, not shown here.)

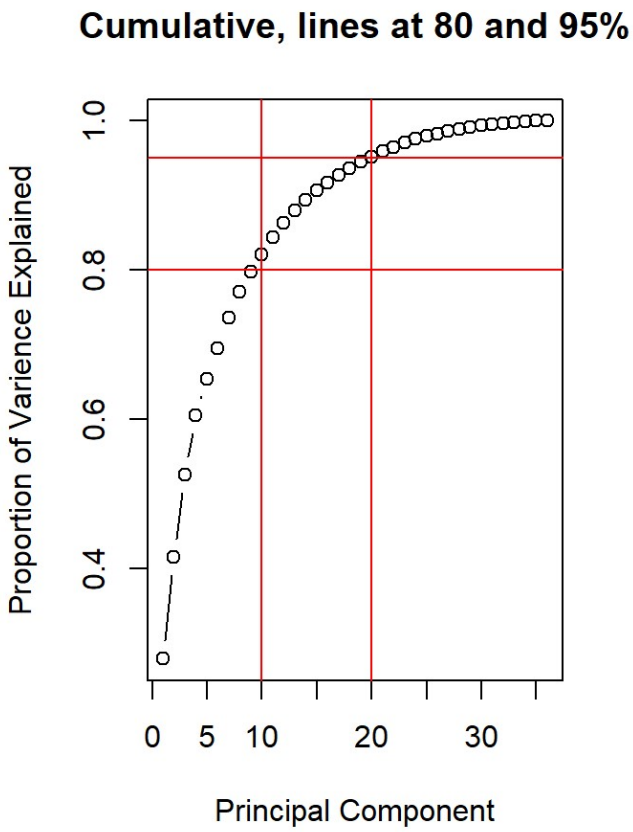
Step 1: Download data into R and make workable dataframes for analysis with all data and maternal diet-based subsets.

```
DGs1.2 <- read.csv("1.2DGs.changed LOQ.csv", head = TRUE, row.names = 1)
AC <- read.csv("AC.formatted.noloq.csv", head = TRUE, row.names = 1)
Cer <- read.csv("Cerimides.changedLOQ.csv", head = TRUE, row.names = 1)
DGs1.3 <- read.csv("CM033018 1-3DGs.changedLOQ.csv", head = TRUE, row.names = 1)
dh <- read.csv("dhCer.changedLOQ.csv", head = TRUE, row.names = 1)
Glu <- read.csv("GluCer.changedLOQ.csv", head = TRUE, row.names = 1)
hex <- read.csv("hexosylCer.changedLOQ.csv", head = TRUE, row.names = 1)
Lac <- read.csv("LacCer.changedLOQ.csv", head = TRUE, row.names = 1)
mye <- read.csv("Sphingomyelins.formatted.noloq.csv", head = TRUE, row.names = 1)
sine <- read.csv("Sphingosine.formatted.noloq.editfordismatrix.csv", header = TRUE, ro
w.names = 1)
TAG <- read.csv("TAG.changedLOQ.csv", head = TRUE, row.names = 1)


#Create a dataframe with all PCA
ALL <- cbind(AC, Cer, DGs1.2, DGs1.3, dh, Glu, hex, Lac, mye, sine, TAG)


#Create a list of of all data structures and the names of each.
plotlist <- list(DGs1.2, DGs1.3, TAG, AC, dh, Glu, hex, Lac, sine, Cer, mye, ALL)
plotname <- c("DGs1.2", "DGs1.3", "TAG", "AC", "dh", "Glu", "hex", "Lac", "sine", "Ce
r", "mye", "ALL")


#Clean up labels in ALL
Labels <- gsub("..pmol.", "", colnames(ALL))
Labels <- gsub("X", "", Labels)
colnames(ALL) <- Labels

#Create subsets
CTR <- rbind(ALL[1:8,],ALL[14:21,],ALL[30:36,])
HFD <- rbind(ALL[9:13,],ALL[22:29,])
```

Next, we calculate the principal compnents with prcomp (). Scree plots show how each principal component describes the data and how many of these components we should reasonably consider to describe data trends.

```
#Principal component
prin_comp <- prcomp(ALL, scale. = T, center = T)

#Varience calculation
Totalvar <- (prin_comp$sdev)^2
Proportionvar <- Totalvar/sum(Totalvar)

#plotting
par(mfrow=c(1,2))

#scree plot

plot(Proportionvar, xlab = "Principal Component", ylab = "Proportion of Varience Expla
ined", type = "b", main = "Proportion for principal component")

#cumulative scree plot

plot(cumsum(Proportionvar),xlab = "Principal Component", ylab = "Proportion of Varienc
e Explained", type = "b", main = "Cumulative, lines at 80 and 95%")
abline(h=0.95, v=20, col="red")
abline(h=0.80, v=10, col="red")
```

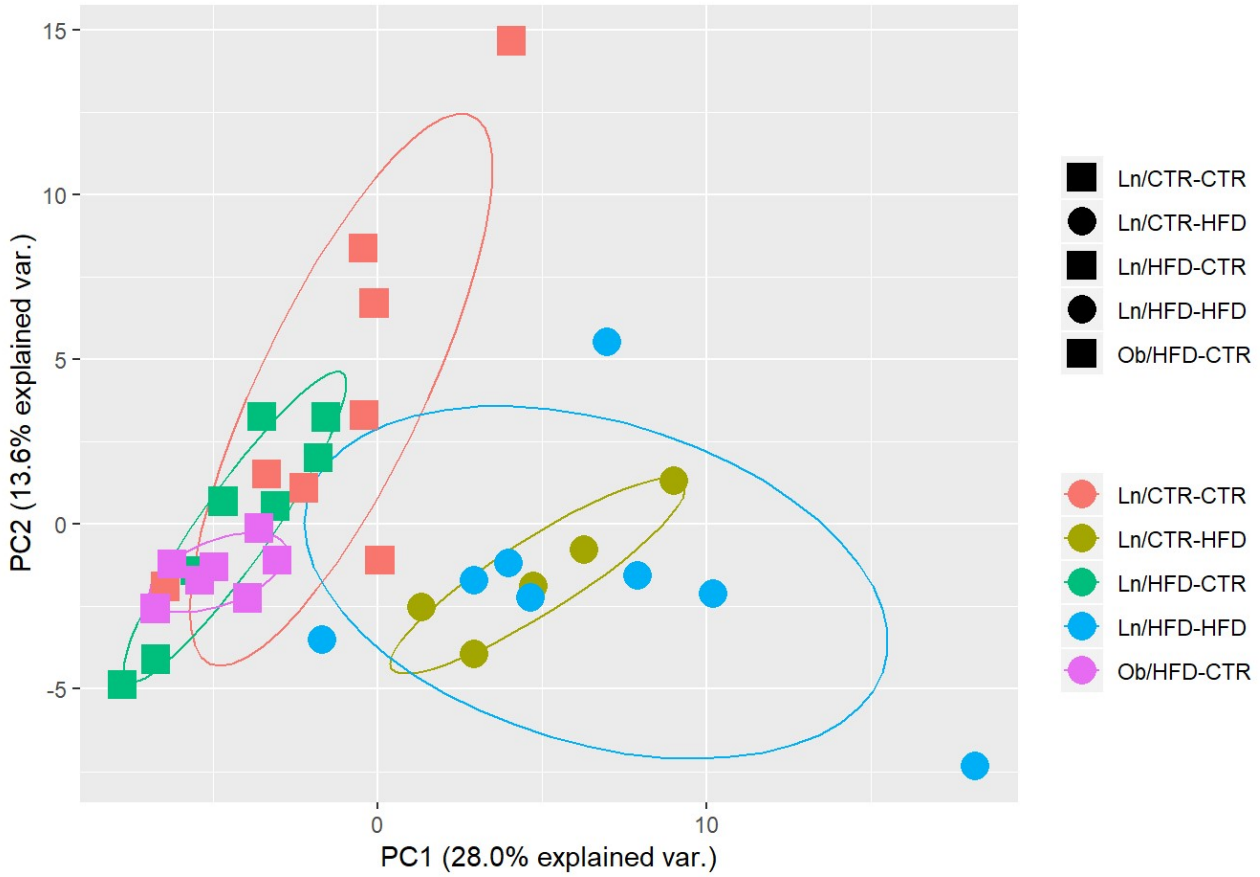




We've elected to use several additional packages to help us visualize our PCA:

```
library(ggpubr) #For arranging biplots
library(devtools) #For arranging biplots
library(ggbiplot) #For arranging all plots

library("FactoMineR") #Specific for PCA visualizing
library("factoextra") #Specific for PCA visualizing
```

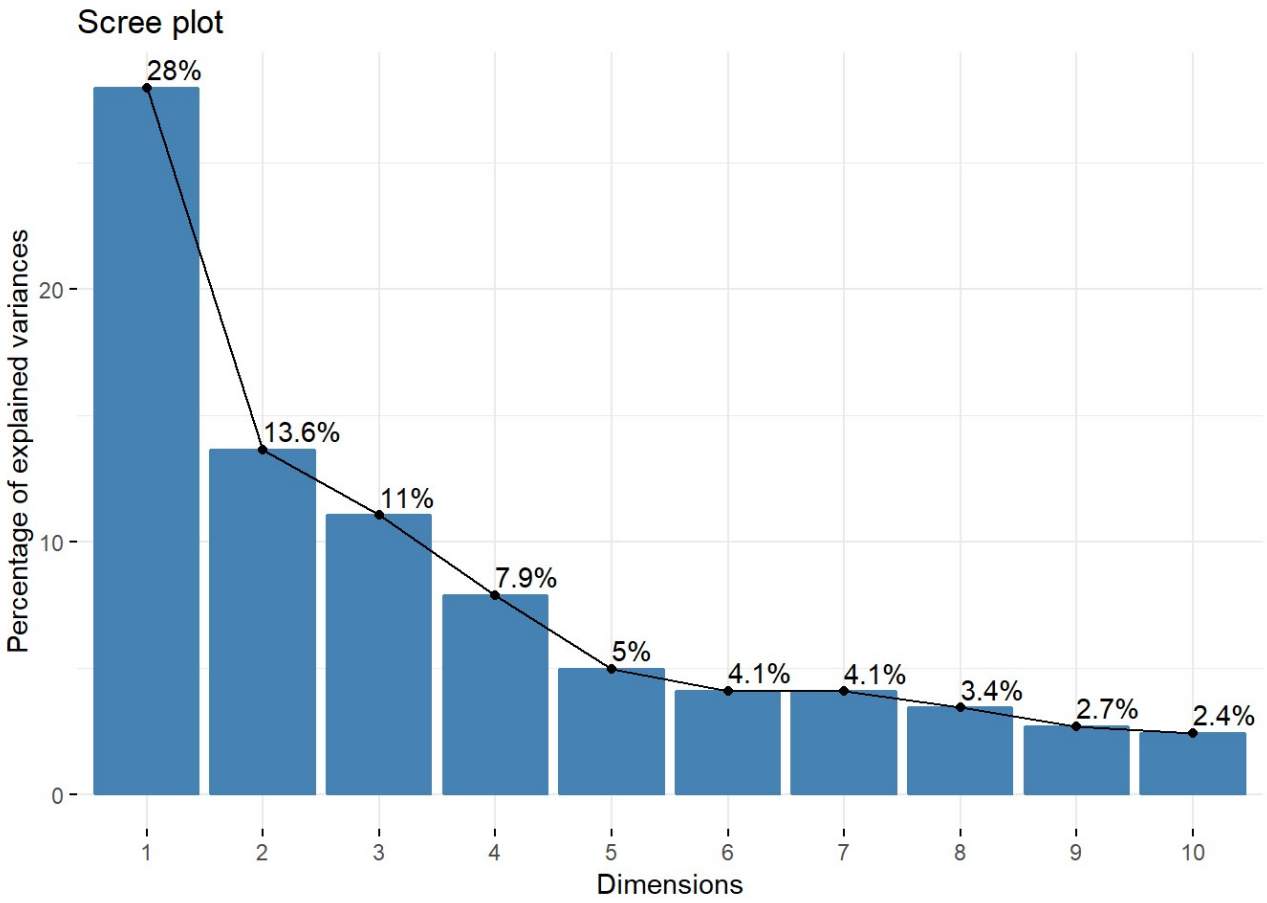
```
#Display PCA; png saving options anf format adjustments are commented out.

###png("PCA1.png", height = 600, width = 800)

#Define sample groups (update with metadata to observe possible batch effects)
groups  <-  c(rep("Ln/CTR-CTR",8),rep("Ln/CTR-HFD",5),rep("Ln/HFD-CTR",8),rep("Ln/HFD-
HFD",8),rep("Ob/HFD-CTR",7))

#Plot with circles
g <- ggbiplot(prin_comp, obs.scale = 1, var.scale = 1, groups = groups, ellipse = TRU
E, circle = TRUE, var.axes = FALSE, varname.size = 0) +

  geom_point(aes(colour=groups, shape = groups), size = 5)+ #Plot points

  scale_shape_manual(name= "", values = c(15,16,15,16,15))+ #Manually adjested shape b
ased on Juvinille diet update with metadata to observe possible batch effects)
  scale_color_manual(name="", values=c("#F8766D", "#A3A500", "#00BF7D", "#00B0F6", "#E
76BF3"))+ #Experimental group colors
  guides(shape = guide_legend(order = 1), color = guide_legend(order = 2)) # Order leg
ends


#g <- g + theme_light() +
#          theme(panel.grid = element_line(colour = "white"),
#               legend.direction = 'vertical',
#                legend.position = c(0.85, 0.75),
#               axis.text.x = element_text(size = 26), axis.text.y = element_text(size
= 26),
#                axis.title.x = element_text(size = 30), axis.title.y = element_text(si
ze = 30),
#               legend.text = element_text(size = 24), legend.title = element_text(siz
e = 0))


print(g) #Show plot
```



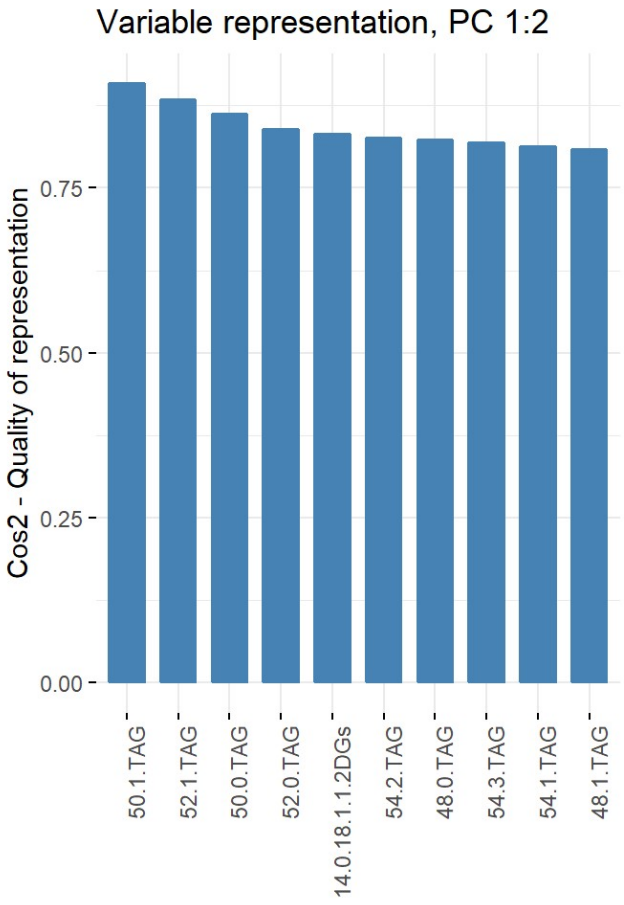

```
###dev.off()
```

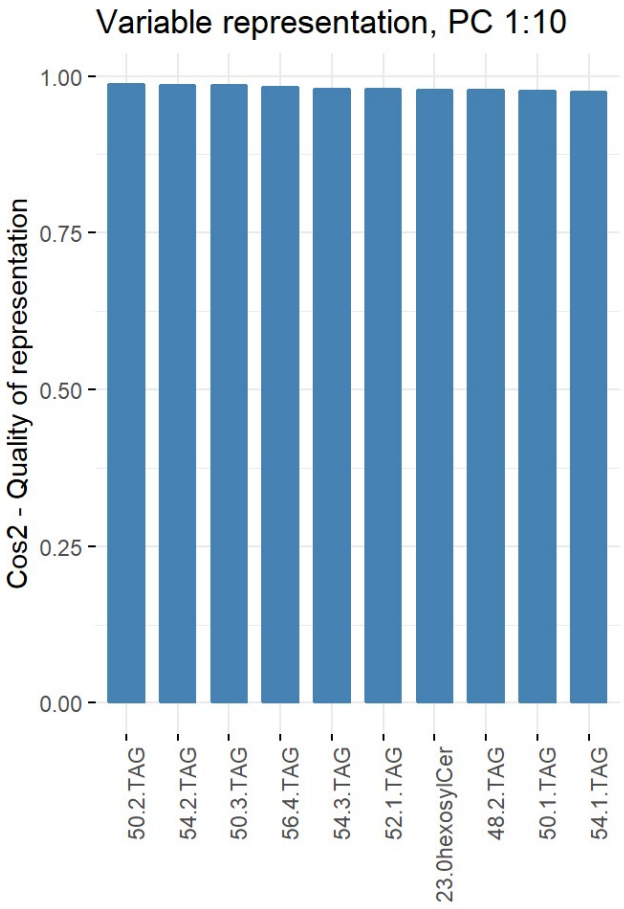
Using these packages, we can recreate our scree plot and investigate the variance explained by the calculated principal components and which lipids are contributing the most to each component.

```
PCA_36 <- PCA(ALL, scale.unit = TRUE, ncp = 36, graph = FALSE) #PCA used in the PCA sp
ecific package

#scree plot
fviz_eig(PCA_36, addlabels = TRUE)
```

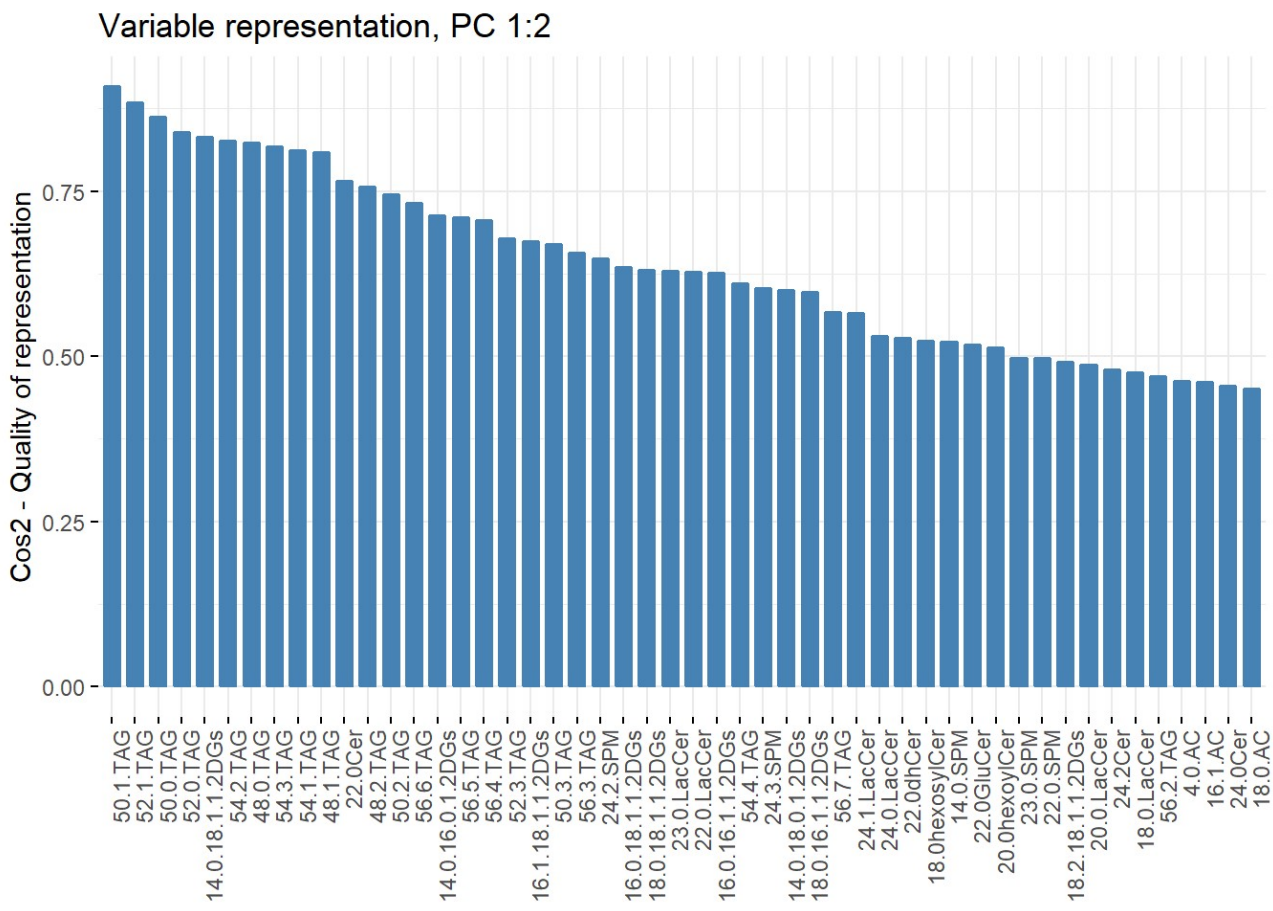



```
#Variable represation for the first two prinicpal components (as used in the PCA), and out to 10 principal components (Noted to account for 80% of variation in the data based on earlier scree plots.)

qual1 <- fviz_cos2(PCA_36, choice = "var", axes = 1:2, top = 10, xtickslab.rt = 90)+ labs(title = "Variable representation, PC 1:2")
qual2 <- fviz_cos2(PCA_36, choice = "var", axes = 1:10, top = 10, xtickslab.rt = 90)+ labs(title = "Variable representation, PC 1:10")
qual3 <- fviz_cos2(PCA_36, choice = "var", axes = 1:2, top = 50, xtickslab.rt = 90)+ labs(title = "Variable representation, PC 1:2")
qual4 <- fviz_cos2(PCA_36, choice = "var", axes = 1:10, top = 50, xtickslab.rt = 90)+ labs(title = "Variable representation, PC 1:10")


ggarrange(qual1, qual2, ncol = 2, nrow = 1)
```

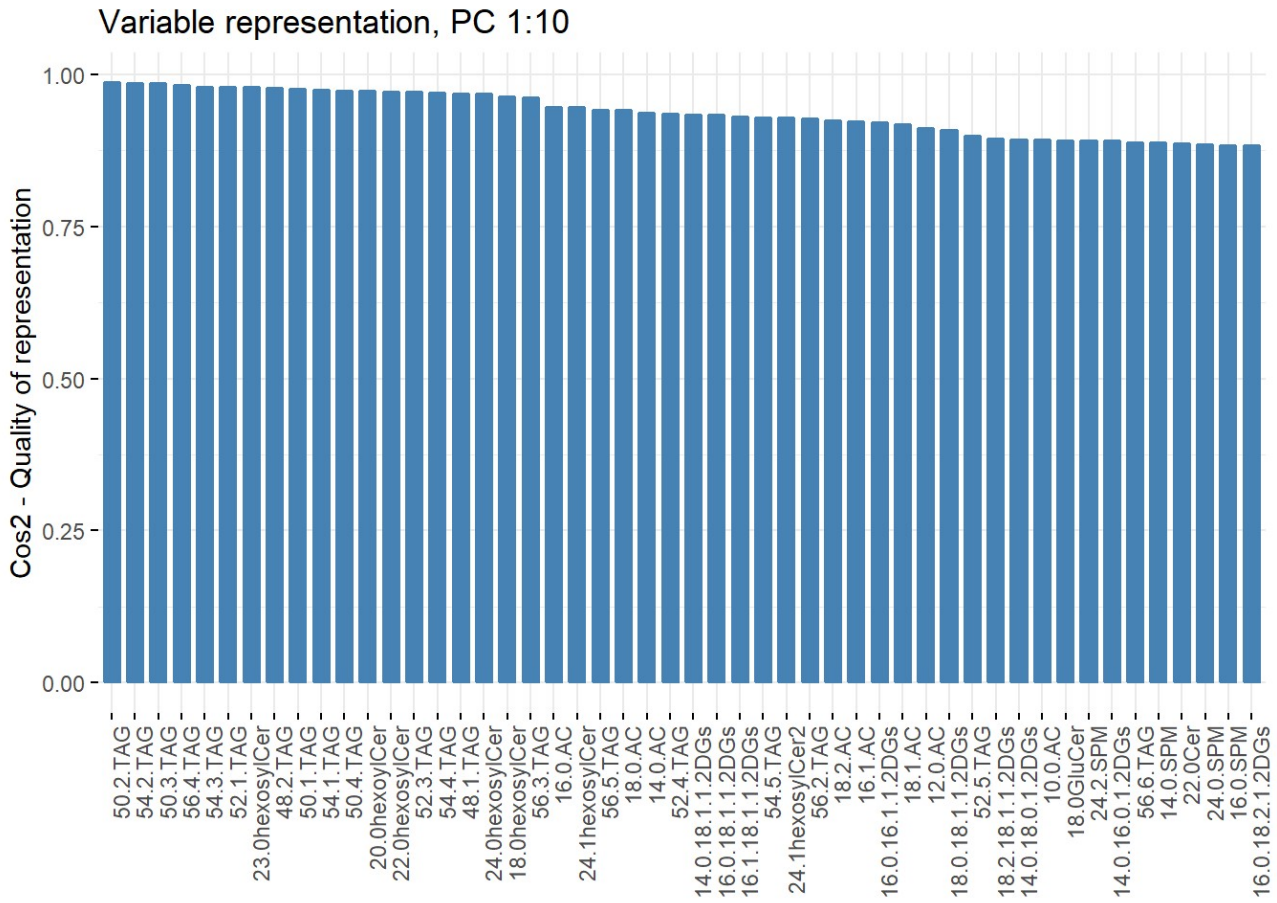

Variable representation, PC 1:2
Cos2 - Quality of representation
0.00
0.25
0.50
0.75
50.1.TAG
52.1.TAG
50.0.TAG
52.0.TAG
14.0.18.1.1.2DGs
54.2.TAG
48.0.TAG
54.3.TAG
54.1.TAG
48.1.TAG


Variable representation, PC 1:10
Cos2 - Quality of representation
0.00
0.25
0.50
0.75
1.00
50.2.TAG
54.2.TAG
50.3.TAG
56.4.TAG
54.3.TAG
52.1.TAG
23.0hexosylCer
48.2.TAG
50.1.TAG
54.1.TAG

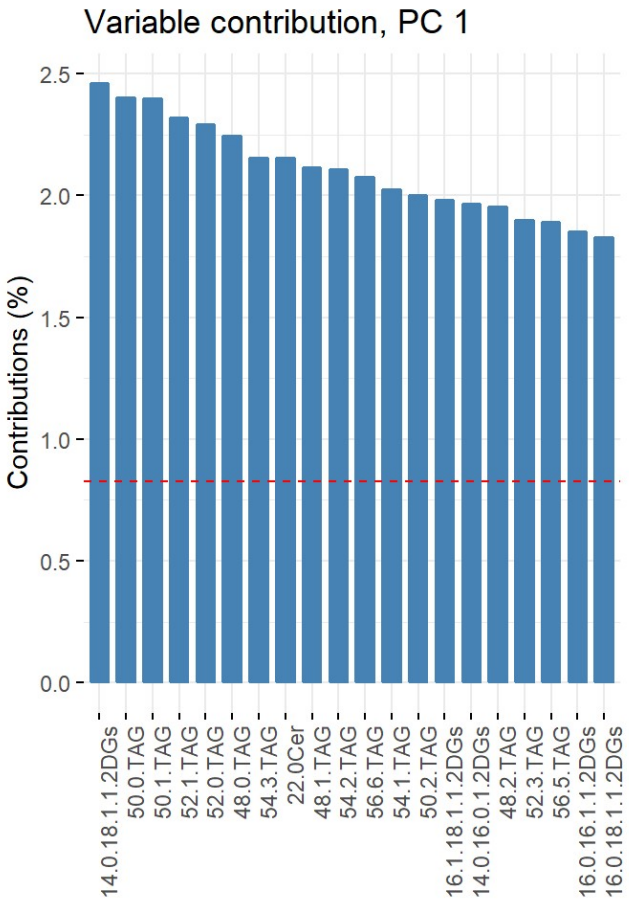
```
qual3
```

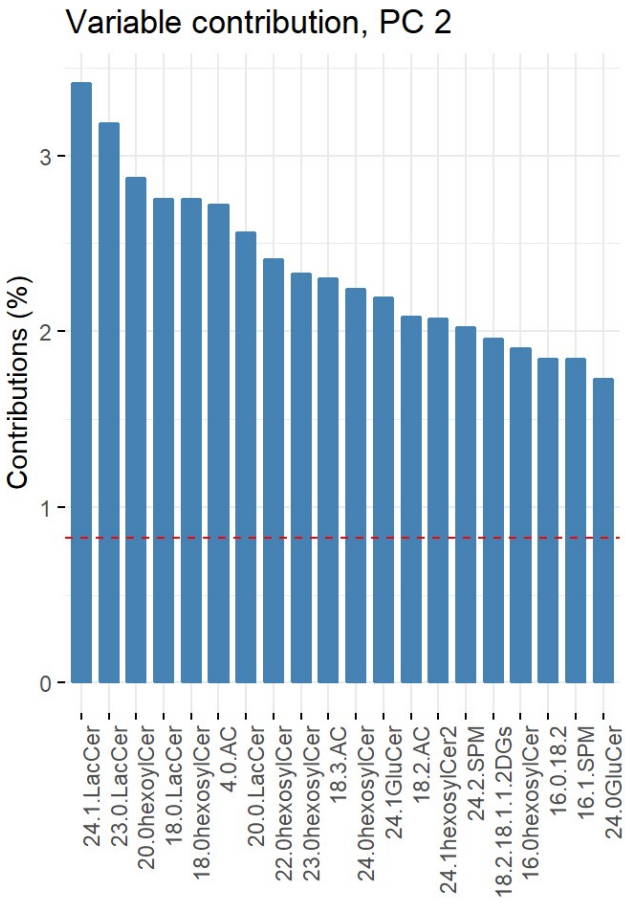

Variable representation, PC 1:2
Cos2 - Quality of representation
0.00
0.25
0.50
0.75
50.1.TAG
52.1.TAG
50.0.TAG
52.0.TAG
14.0.18.1.1.2DGs
54.2.TAG
48.0.TAG
54.3.TAG
54.1.TAG
48.1.TAG
22.0Cer
48.2.TAG
50.2.TAG
56.6.TAG
14.0.16.0.1.2DGs
56.5.TAG
56.4.TAG
52.3.TAG
16.1.18.1.1.2DGs
50.3.TAG
56.3.TAG
24.2.SPM
16.0.18.1.1.2DGs
18.0.18.1.1.2DGs
23.0.LacCer
22.0.LacCer
16.0.16.1.1.2DGs
54.4.TAG
24.3.SPM
14.0.18.0.1.2DGs
18.0.16.1.1.2DGs
56.7.TAG
24.1.LacCer
24.0.LacCer
22.0dhCer
18.0hexosylCer
14.0.SPM
22.0GluCer
20.0hexoylCer
23.0.SPM
22.0.SPM
18.2.18.1.1.2DGs
20.0.LacCer
24.2Cer
18.0.LacCer
56.2.TAG
4.0.AC
16.1.AC
24.0Cer
18.0.AC

```
qual4
```

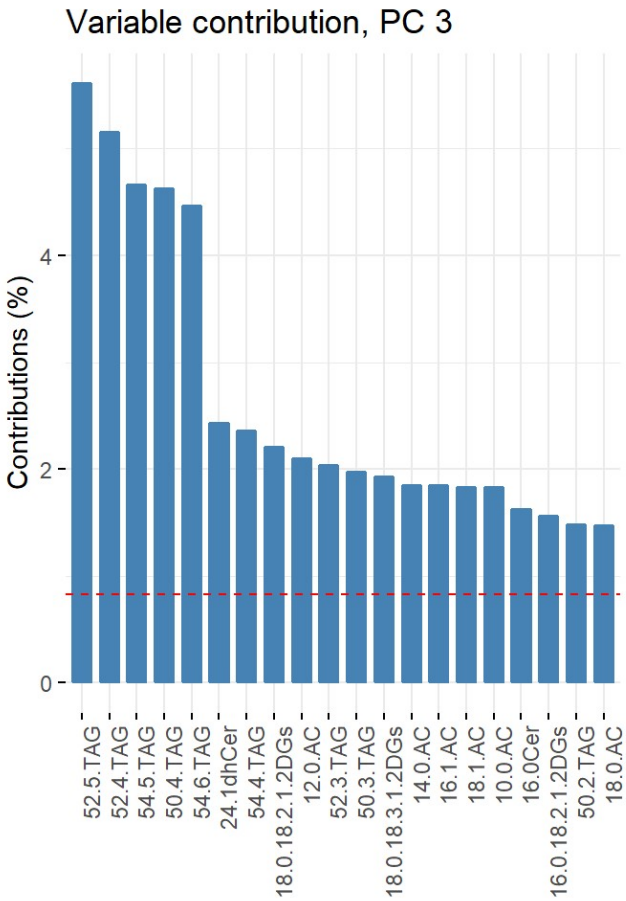


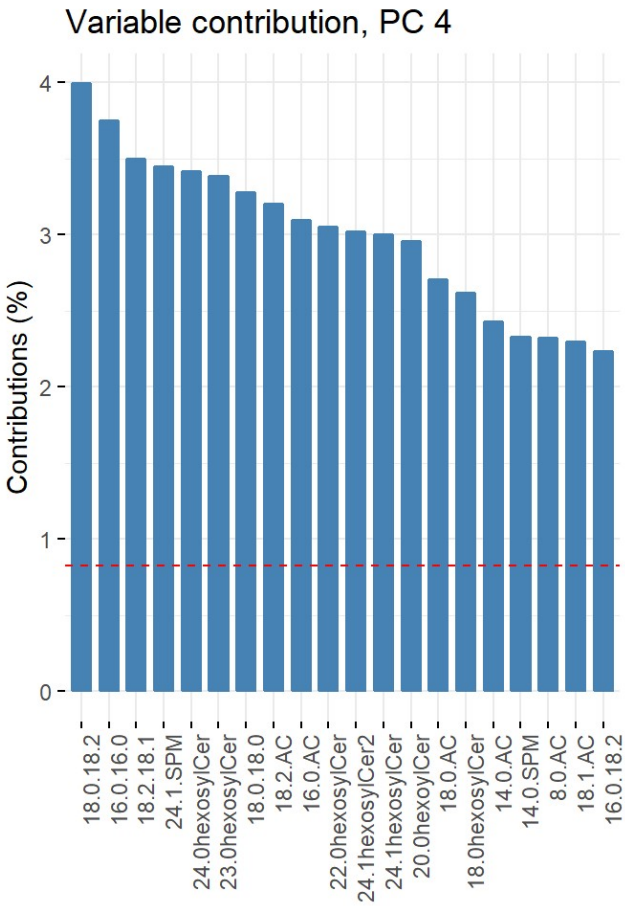

```
#Lipid contribution to the first four prinicpal components (as used in the PCA), and o
ut to 10 and 20 principal components (Noted to account for 80%, 95% of variation in th
e data based on earlier scree plots.)

con1 <- fviz_contrib(PCA_36, choice = "var", axes = 1, top = 20, xtickslab.rt = 90)+ l
abs(title = "Variable contribution, PC 1")
con2 <- fviz_contrib(PCA_36, choice = "var", axes = 2, top = 20, xtickslab.rt = 90)+ l
abs(title = "Variable contribution, PC 2")
con3 <- fviz_contrib(PCA_36, choice = "var", axes = 3, top = 20, xtickslab.rt = 90)+ l
abs(title = "Variable contribution, PC 3")
con4 <- fviz_contrib(PCA_36, choice = "var", axes = 4, top = 20, xtickslab.rt = 90)+ l
abs(title = "Variable contribution, PC 4")

ggarrange(con1, con2, ncol = 2, nrow = 1)
```



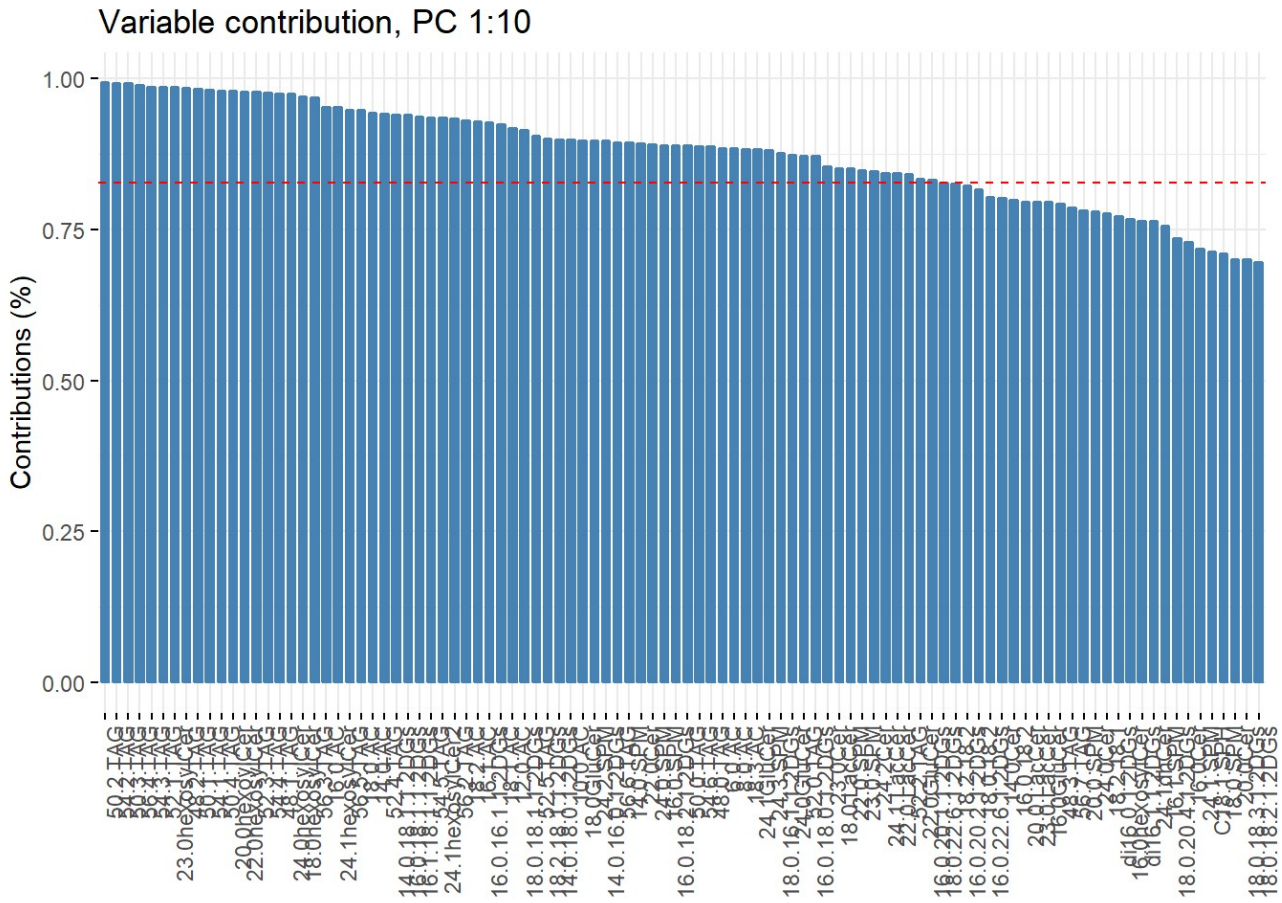



```
ggarrange(con3,con4, ncol = 2, nrow = 1)
```

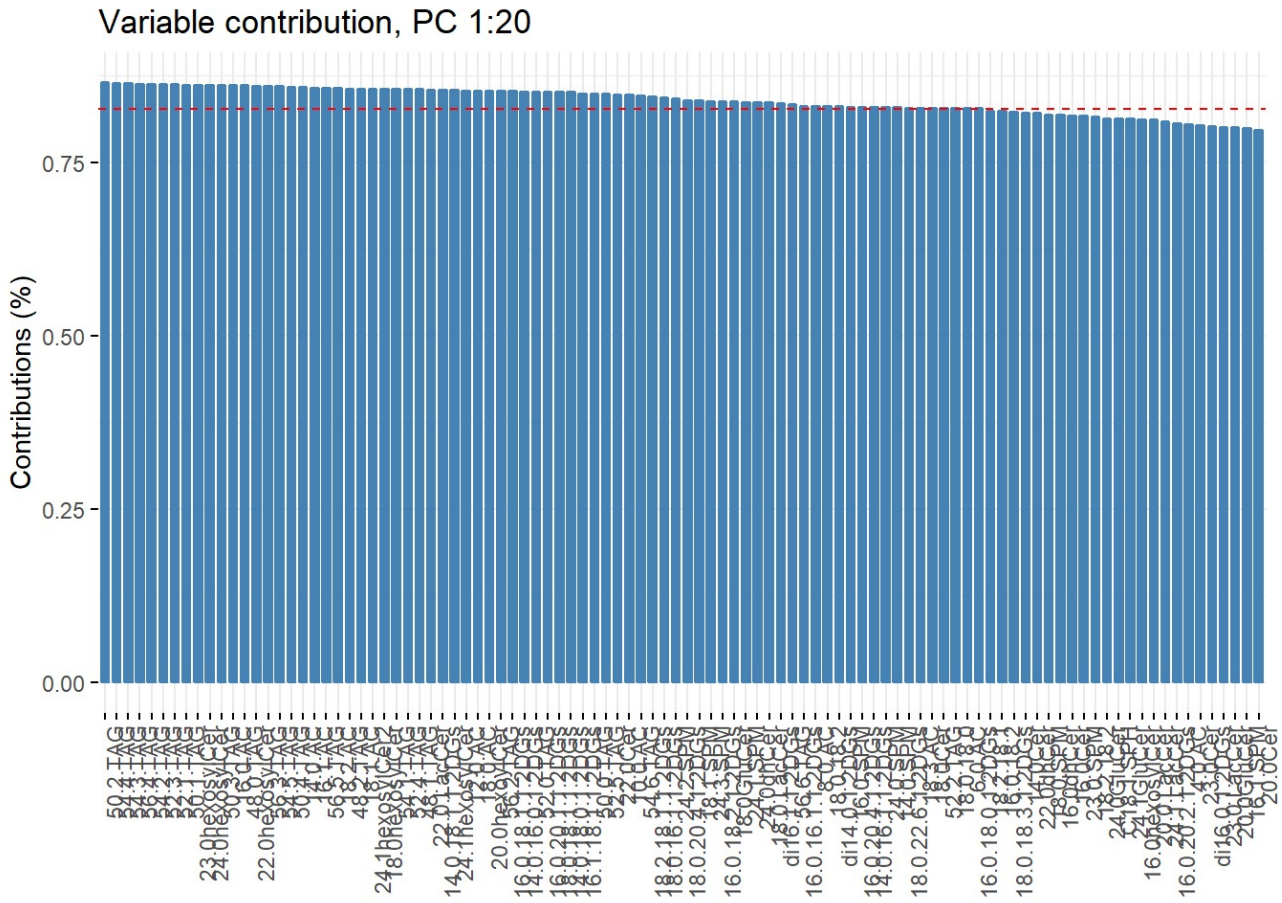




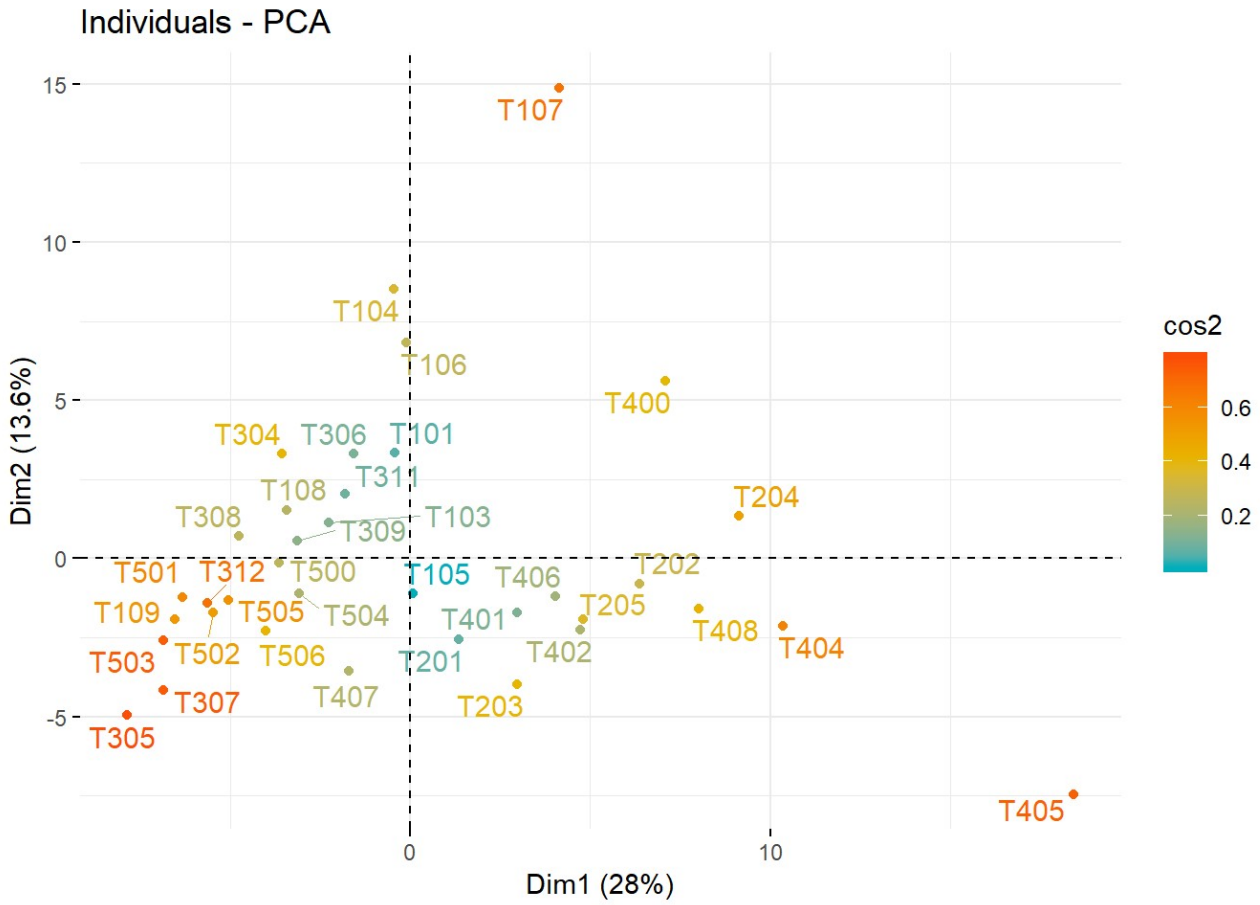

```
fviz_contrib(PCA_36, choice = "var", axes = 1:10, top = 100, xtickslab.rt = 90)+ labs
(title = "Variable contribution, PC 1:10")
```



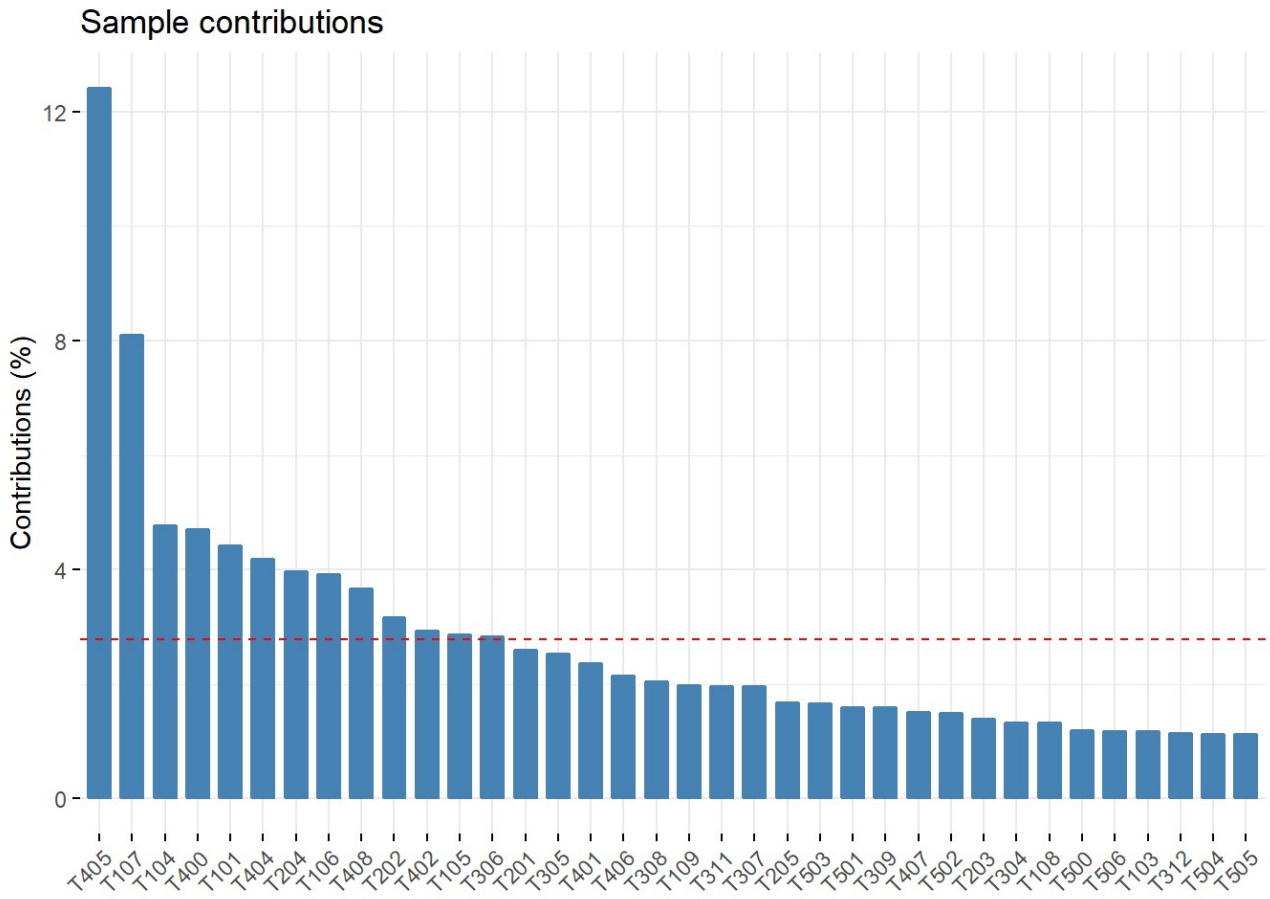

```
fviz_contrib(PCA_36, choice = "var", axes = 1:20, top = 100, xtickslab.rt = 90)+ labs
(title = "Variable contribution, PC 1:20")
```




```
#Redlines represent overall mean

#Labeled PCA with sample names and contribution to PCA from lipids in each sample. Not
ice that extreme points have the most significant contributions.

fviz_pca_ind(PCA_36, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping (slow if many points)
             )
```




```
#Associated sample contribution bar graph
fviz_contrib(PCA_36, choice = "ind", axes = 1:35)+ labs(title = "Sample contribution
s")
```




```
#var$cos2: represents the quality of representation for variables on the factor map. I
t's calculated as the squared coordinates: var.cos2 = var.coord * var.coord.
```

Swapping out the All dataframe with the subsets of HFD and CTR, isolated trends within maternal diet can be discerned as well.