GXE Mouse Data Walkthrough

Jameson Hinkle

April 13, 2020

Contents


```
library(tidyverse)
library(ggpubr)
library("FactoMineR")
library("factoextra")
```

Gene By Environment Interaction can give clues to personalized diet management.

So to begin, because we measured these mice for multiple phenotypes through the experimental timeline, these data are inherently multi-variate. This means that there is likely signal in the data that is nuanced that has interactions we can't see just by doing a simple regression or categorical comparison. Prior to performing any sort of modeling, it would be nice to see what variables contribute to the variance seen in the data so we can make informed decisions about what variables (if any) underly the GXE interaction. One way to look at what variables(Phenotypes) contribute to this variance is by dimensionality reduction. One of the oldest and most used dimensionality reduction methods is a Principle Component Analysis (PCA, click link). This method looks to rotate the data to look for orthogonals in the data that show different contributions by different variables (Phenotypes). In our case, we will use PCA to reduce the number of dimensions down to just a few, so we can see what the differences are among the variables that contribute to the dataset. Therefore, for this experiment we will look if there is a GXE interaction via PCA bi-plots (looking for dimensional variance in Strain but not Environment (Diet)), Identify variables that contribute to that variance (e.g., Insulin, Glucose),Look to see if there are statistical differences in those data by strain to see differential GXE outcomes.

PCA of measured Phenotypes

```
#read in data
dat <- read_csv("MouseMasterDate_GXE.csv")
#get rid of average data we do not need
dat <- dat %>% select(-BW_Average, -Food_Avg, -Food_Mass)
# select columns to start setting up percent difference in weight
weight <- dat %>% select(`Mouse Cage`, `Mouse Number`, Strain, Diet, Week, PercBWincrease)
View(weight)
# take week 0 for weight to make a baseline
week0 <- weight %>% filter(Week == 0) %>% mutate(W0_mean = PercBWincrease) %>% select(W0_mean, `Mouse Nu
View(week0)
```

```
#calculate percent difference in weight gain
weight1 <- weight %>% left_join(week0, by = c("Mouse Number", "Strain", "Diet")) %>% mutate(percdiff = a
View(weight1)
#check length of PercBWincrease to be sure the vector is the same length to replace raw BW data with pe
length(dat$PercBWincrease)
```

```
## [1] 1694
```

```
length(weight1$percdiff)
```

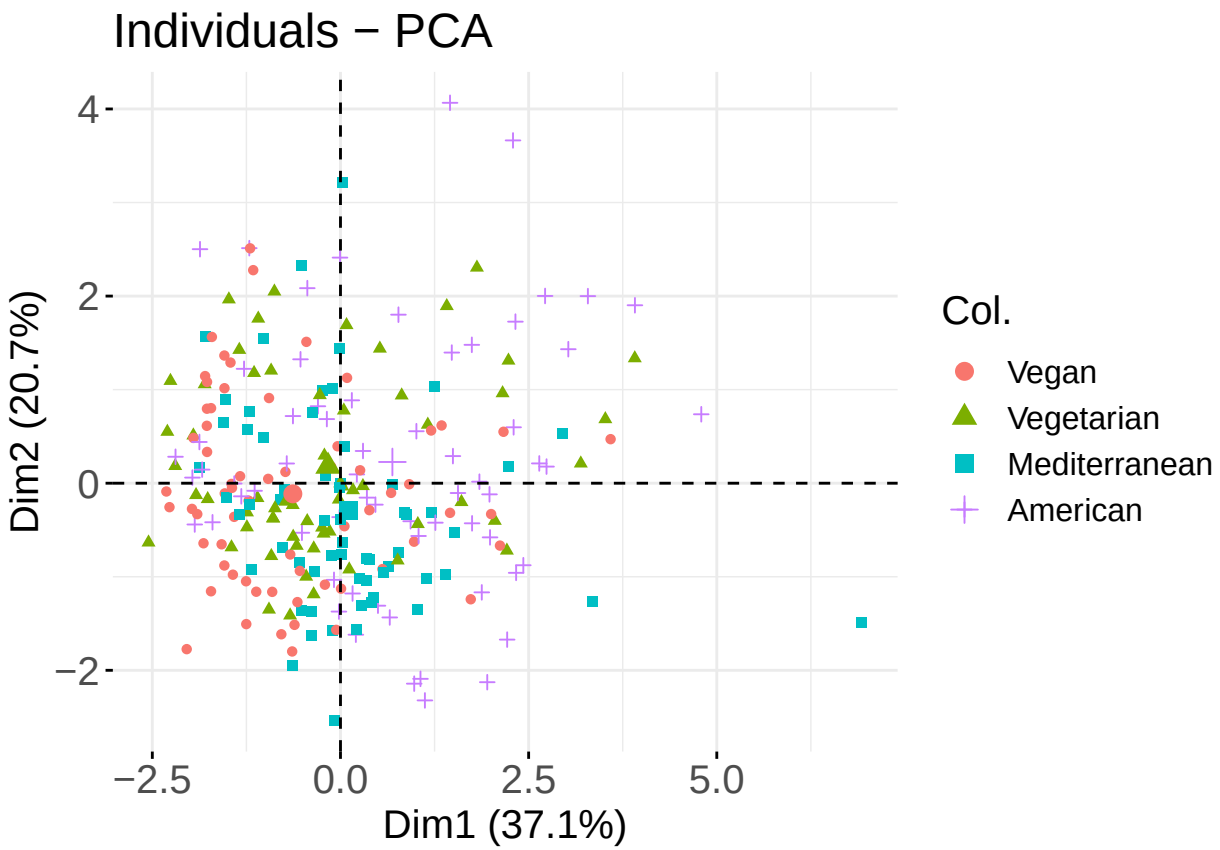
```
## [1] 1694
```

```
dat$PercBWincrease <- weight1$percdiff
week4 <- dat %>% filter(Week == 4)
 week8 <- dat %>% filter(Week == 8)
# create PCA data by binding week 4 and 8 data
# PCA for 2D plotting
PCA_dat_na <- week4 %>% bind_rows(week8)
# select all diets except for Chow for PCA
PCA_dat_na <- PCA_dat_na %>% filter(Diet != "Chow")
# create active numeric columns for PCA
# 2D plotting
active_na <- PCA_dat_na %>% select(PercBWincrease, Ingest_bymouse, Glucose, Trig, NEFA, Insulin)
res.pca <- PCA(active_na, scale.unit = TRUE, graph = FALSE)
```

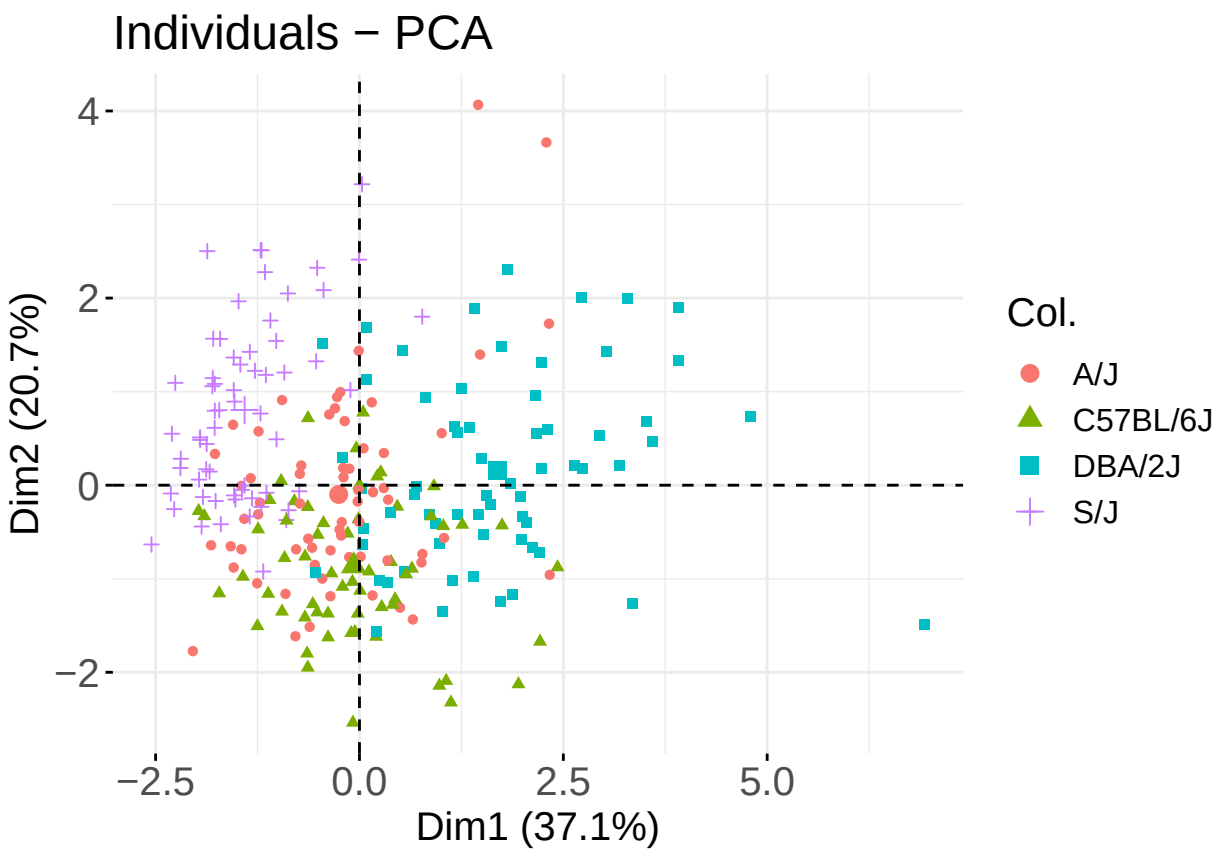
PCA Bi-plot of data

Here we can see our aforementioned PCA bi-plots bi Diet. This is a "bird's eye" look at what the variance in these data look like. I have PC1 verus 2 on the left, and PC2 versus 3 on the right. If there COULD be a GXE interaction, the best thing we could see is a lot of overlap in the variance among diets. The diets could be separated but it would be harder for the data to covary in this way. It is nice that we see a pretty good overlap of the diets here. If there is more variation among the strains, we will be able to begin to think that there is indeed a GXE interaction.

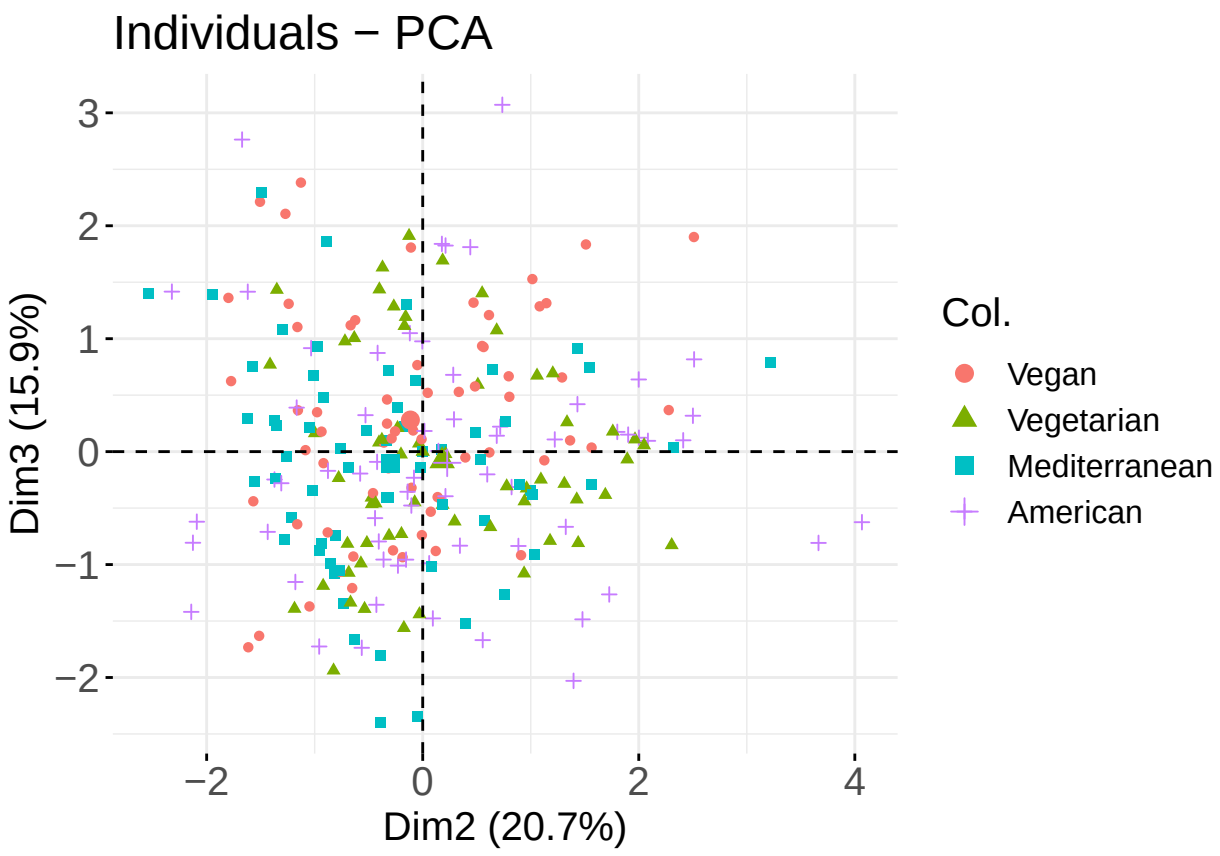
```
# PCA biplot by diet, and strain
# also order diet appropriately
PCA_dat_na$Diet <- factor(PCA_dat_na$Diet, levels = c("Vegan", "Vegetarian", "Mediterranean", "American
fviz_pca_ind(res.pca, geom.ind = "point", col.ind = PCA_dat_na$Diet) +
  theme(text = element_text(size = 15), axis.title = element_text(size = 15), axis.text = element_text(s
```



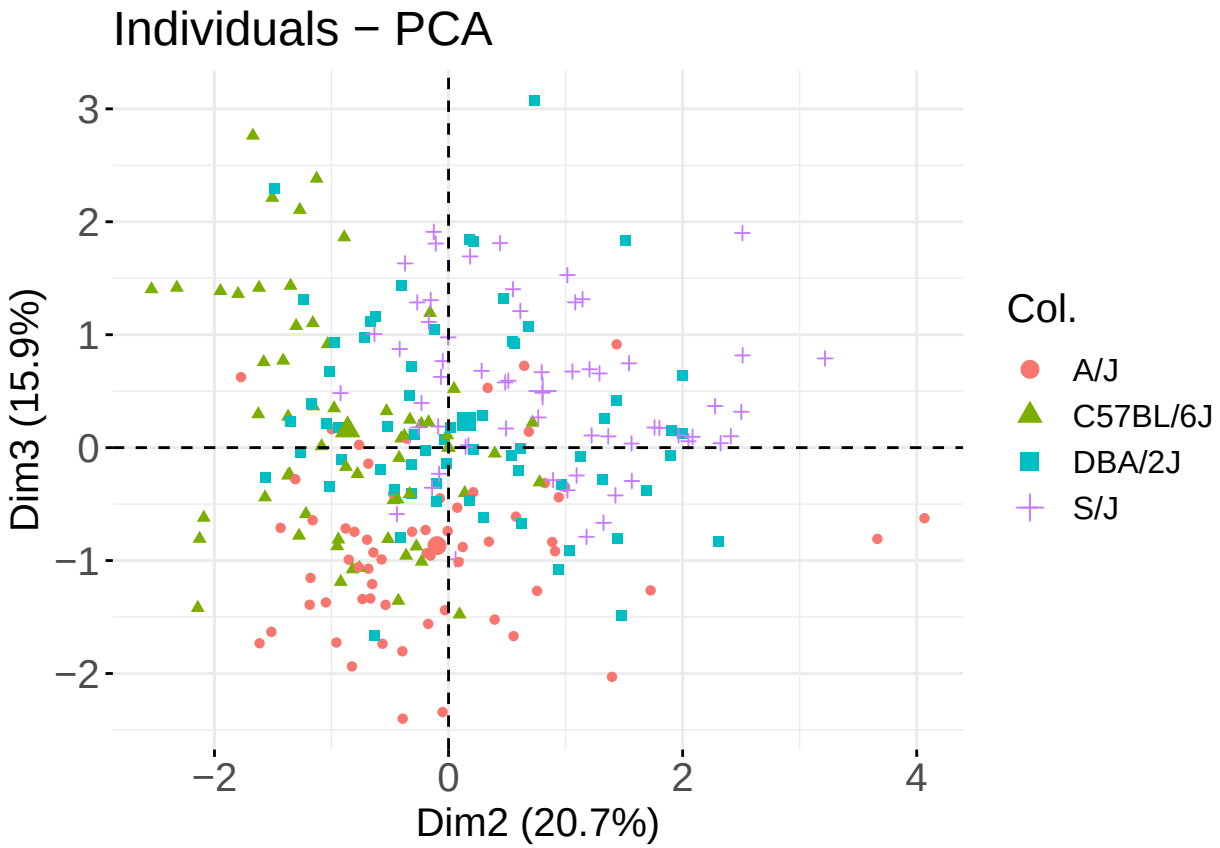

```
fviz_pca_ind(res.pca, geom.ind = "point", col.ind = PCA_dat_na$Strain) +
  theme(text = element_text(size = 15), axis.title = element_text(size = 15), axis.text = element_text(s
```




```
fviz_pca_ind(res.pca, axes = c(2, 3), geom.ind = "point", col.ind = PCA_dat_na$Diet) +
  theme(text = element_text(size = 15), axis.title = element_text(size = 15), axis.text = element_text(s
```




```
fviz_pca_ind(res.pca, axes = c(2, 3), geom.ind = "point", col.ind = PCA_dat_na$Strain) +
  theme(text = element_text(size = 15), axis.title = element_text(size = 15), axis.text = element_text(s
```

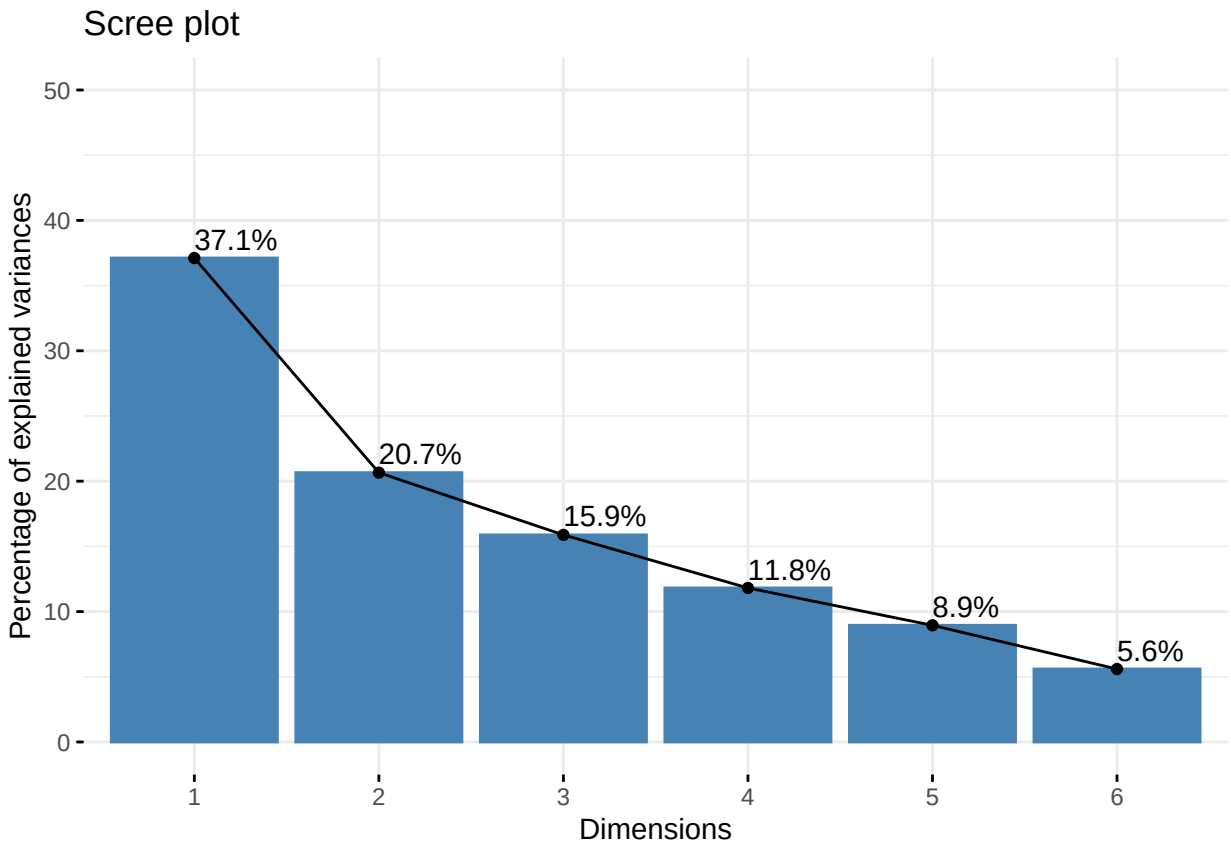


Here we can see PCA bi-plots but this time it is by our mouse strains. Again, at this point if we're going to see GXE interactions, there should be different variances among strain. Indeed, we see this especially in the biplot on the left with PC1/2. We can see that in both bi-plots for example the DBA/2J mice are most separated (most obviously different) amongst the variance space on both the biplots of PC1/2.

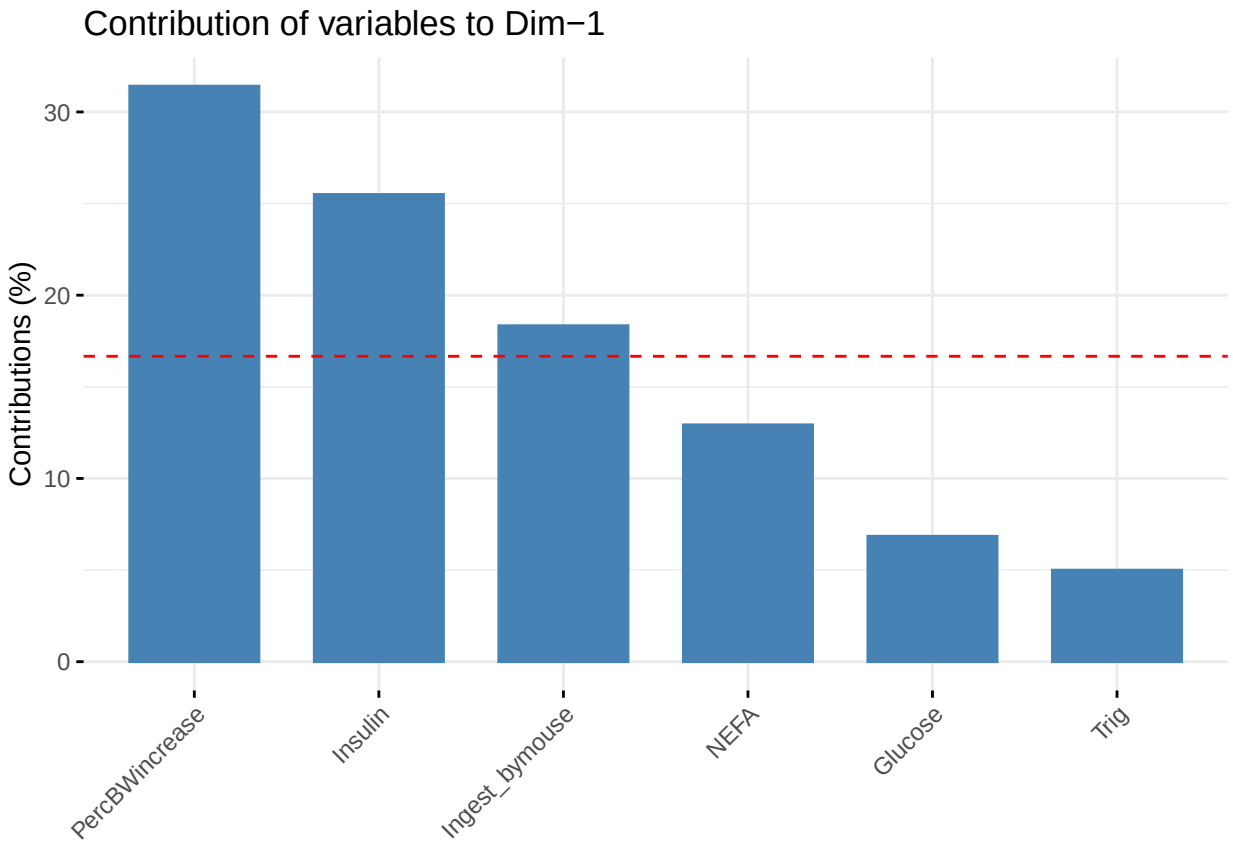
Looking at Variables Contributing to each PC for Further Analysis.

Ok, now we have seen with our "bird's eye" view via biplot that a GXE interaction is possible but, what percentage of the data is explained by these PCs? If you look on the left we have what's called a scree plot. A scree plot shows the amount of variance explained by each PC. If there was no real difference in variance explained, then all the PCs would contribute equally (in this case ~16.7%). Since we see an initial peak of 37.5% in PC1, and it decreases from there, we can see that the first couple (or maybe 3) PCs could contribute to a lot of variance explained in these data. It should be noted that if something like 90% of the variance was explained in one PC in this case, it would likely not be a realistic outcome because we did not control for that much variance in this case. A smooth curve like this in a scree plot is nice in terms of biology because as we know, biology is "the science of exception". Now that we know that PC1 contributes 37.5%, what are the variables that are contributing to that 37.5%? You can see that in the table to the right. Bolded variable percentages are the variables we will look at to see specific GXE differences, and to possibly generate more specific hypotheses.

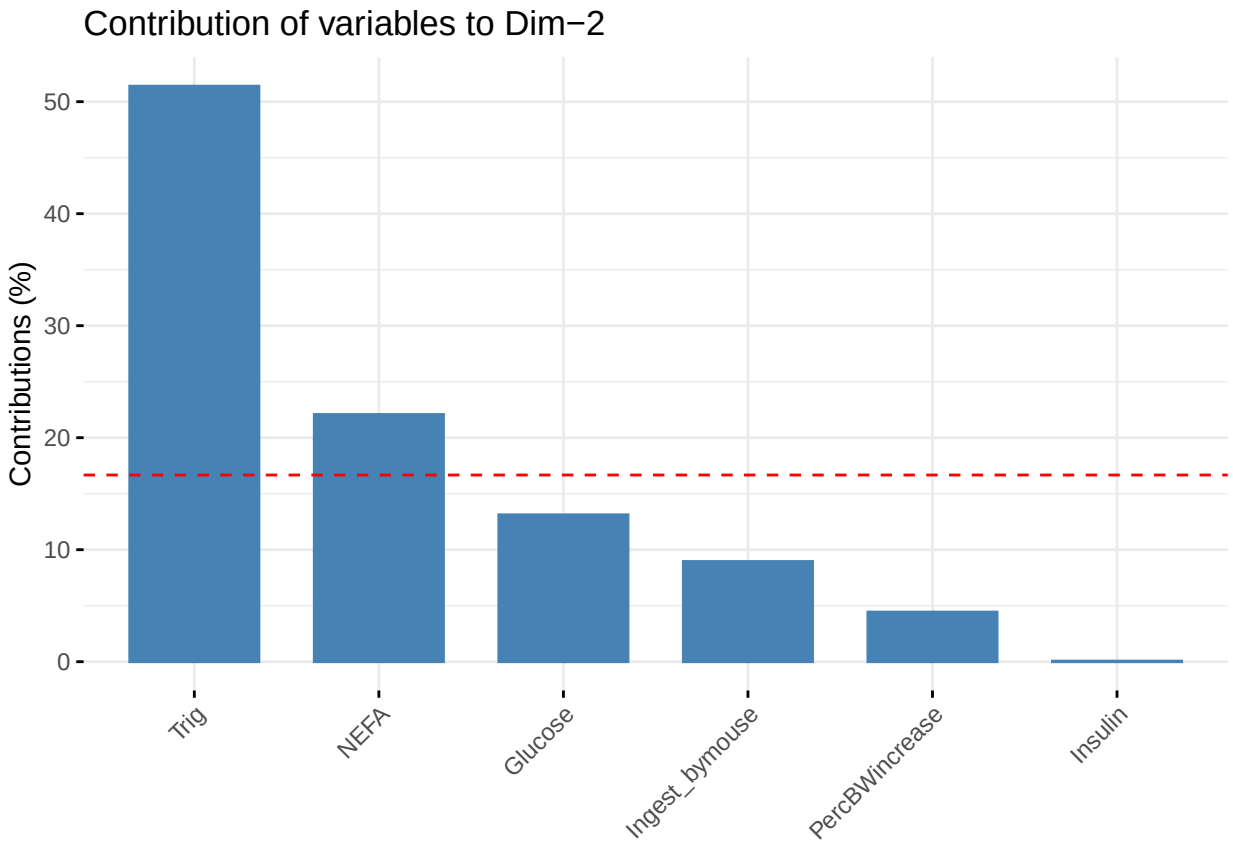
```
# scree plot
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))
```



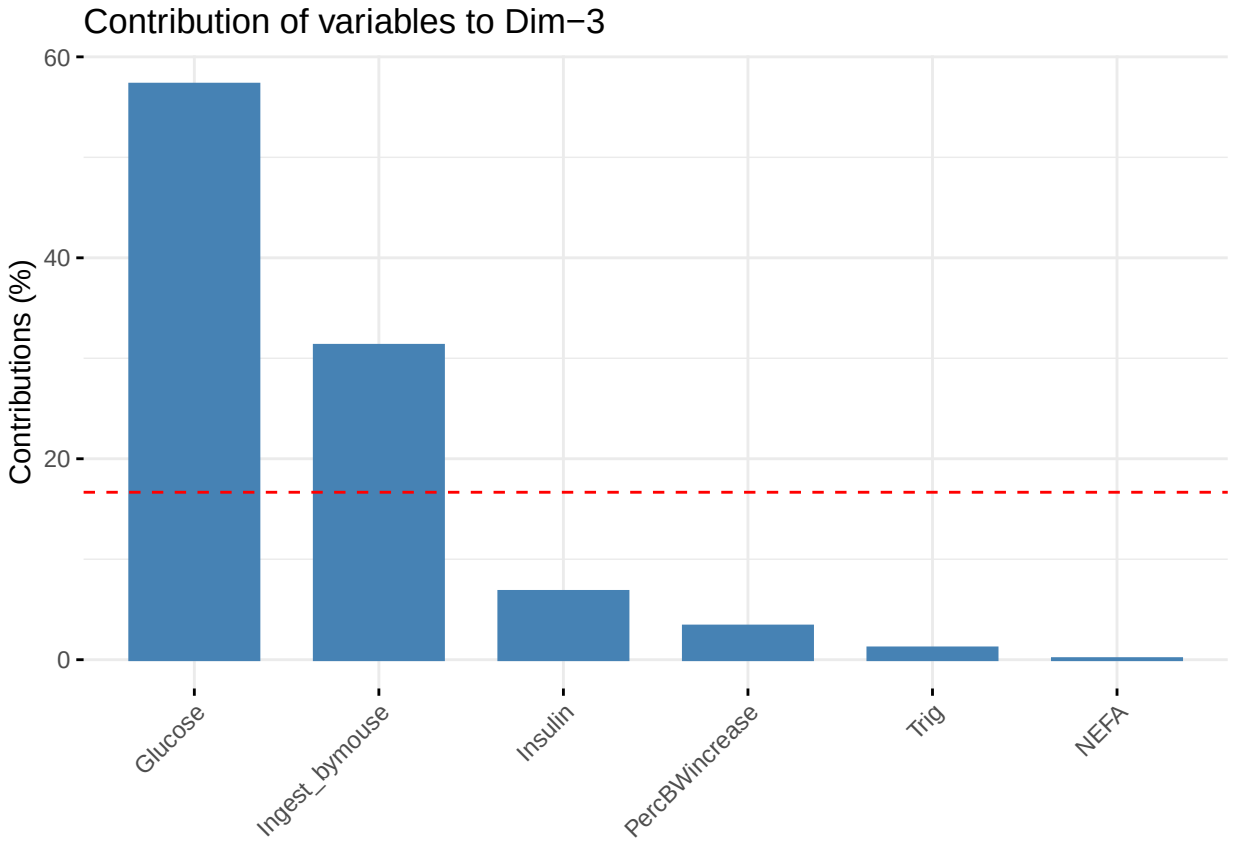

```
# contributions of variables to PC1 and PC2
# extract contributions of variables to respective PCs in a table
var <- get_pca_var(res.pca)
var <- var$contrib
fviz_contrib(res.pca, choice = "var", axes = 1, top = 10)
```




```
fviz_contrib(res.pca, choice = "var", axes = 2, top = 10)
```



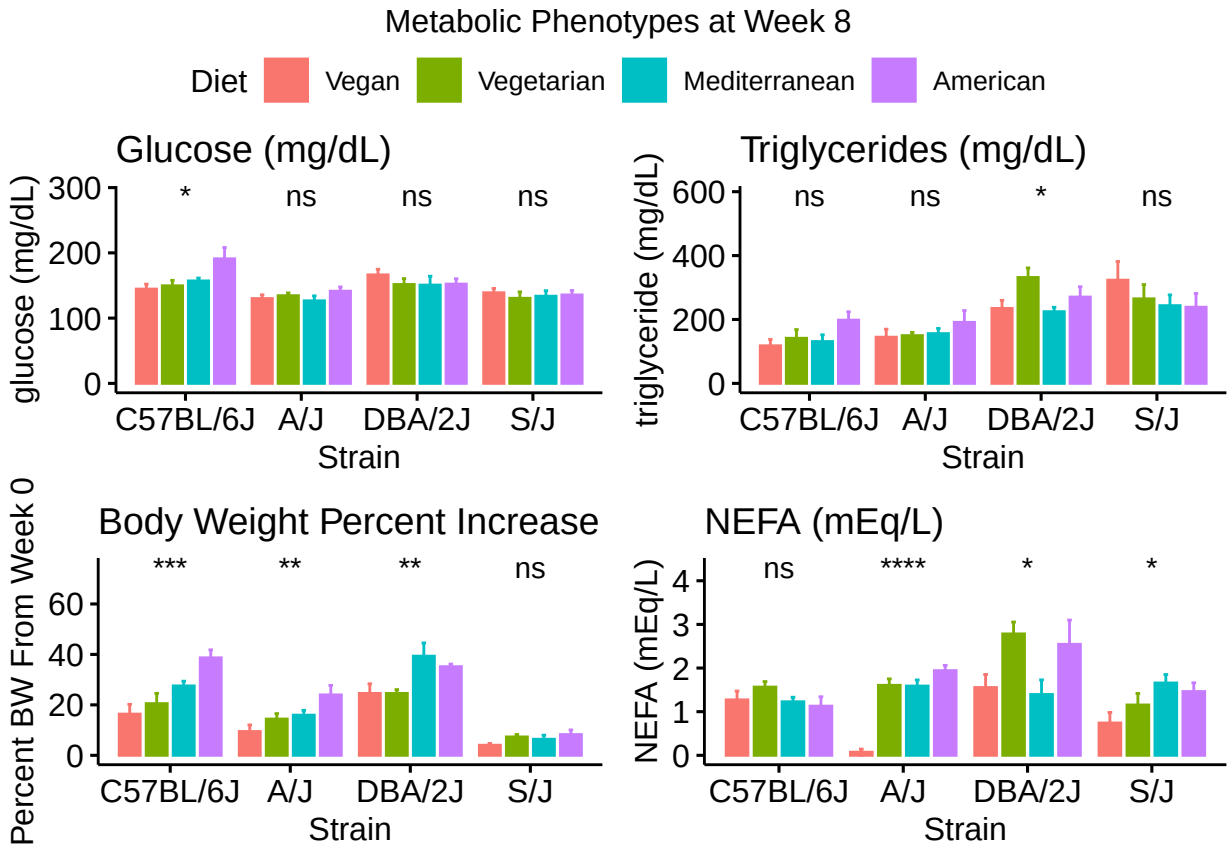

```
fviz_contrib(res.pca, choice = "var", axes = 3, top = 10)
```



Plotting Highest Contributing Variables to look for significant differences in GXE interactions.

Here you can see a barplot of differences of strain by diet tested by a one-way ANOVA. A couple perfect examples of our GXE interactions can be seen in the differences in glucose (C57BL6 is the only strain with a significant difference), and in triglycerides (two strains have significant differences by diet by not all).

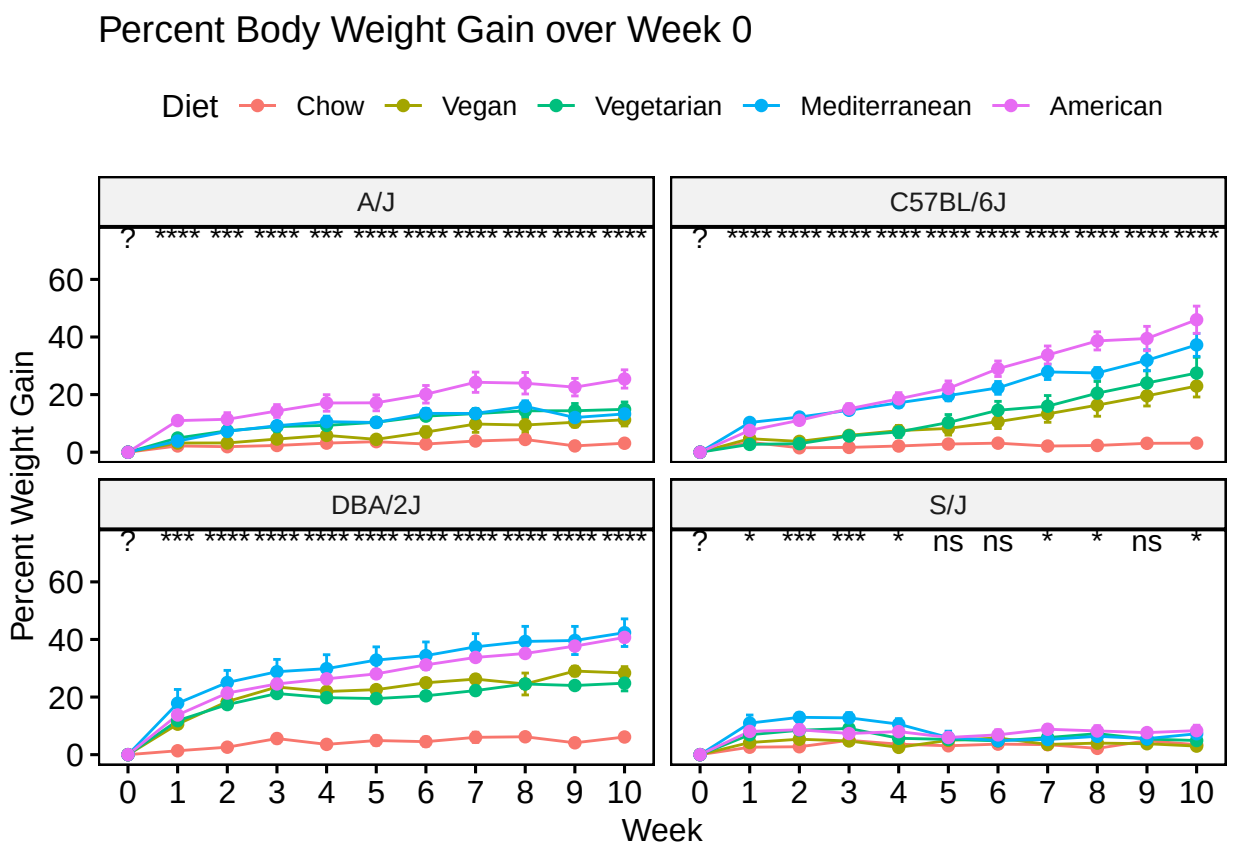
```
# Barplots (Figure 3)
# make diets in order requested
PCA_dat <- PCA_dat_na
fin <- PCA_dat %>% filter(Week == 8)
fin$Diet <- factor(fin$Diet, levels = c("Chow", "Vegan", "Vegetarian", "Mediterranean", "American"))

glucose <- ggbarplot(fin, x = "Strain", y = "Glucose", color = "Diet", fill = "Diet", title = "Glucose
glucose <- glucose + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
trig <- ggbarplot(fin, x = "Strain", y = "Trig", color = "Diet", fill = "Diet", title = "Triglycerides
trig <- trig + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
percdiff <- ggbarplot(fin, x = "Strain", y = "PercBWincrease", color = "Diet", fill = "Diet", title = "
percdiff <- percdiff + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
insulin <- ggbarplot(fin, x = "Strain", y = "Insulin", color = "Diet", fill = "Diet", title = "Insulin"
insulin <- insulin + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
nefa <- ggbarplot(fin, x = "Strain", y = "NEFA", color = "Diet", fill = "Diet", title = "NEFA (mEq/L)",
nefa <- nefa + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
topPCs <- ggarrange(glucose, trig, percdiff, nefa,ncol = 2, nrow =2, common.legend = TRUE)
annotate_figure(topPCs, top = text_grob("Metabolic Phenotypes at Week 8"))
```



Looking at Percent Weight Gain Over Week 0 as a GXE Interaction

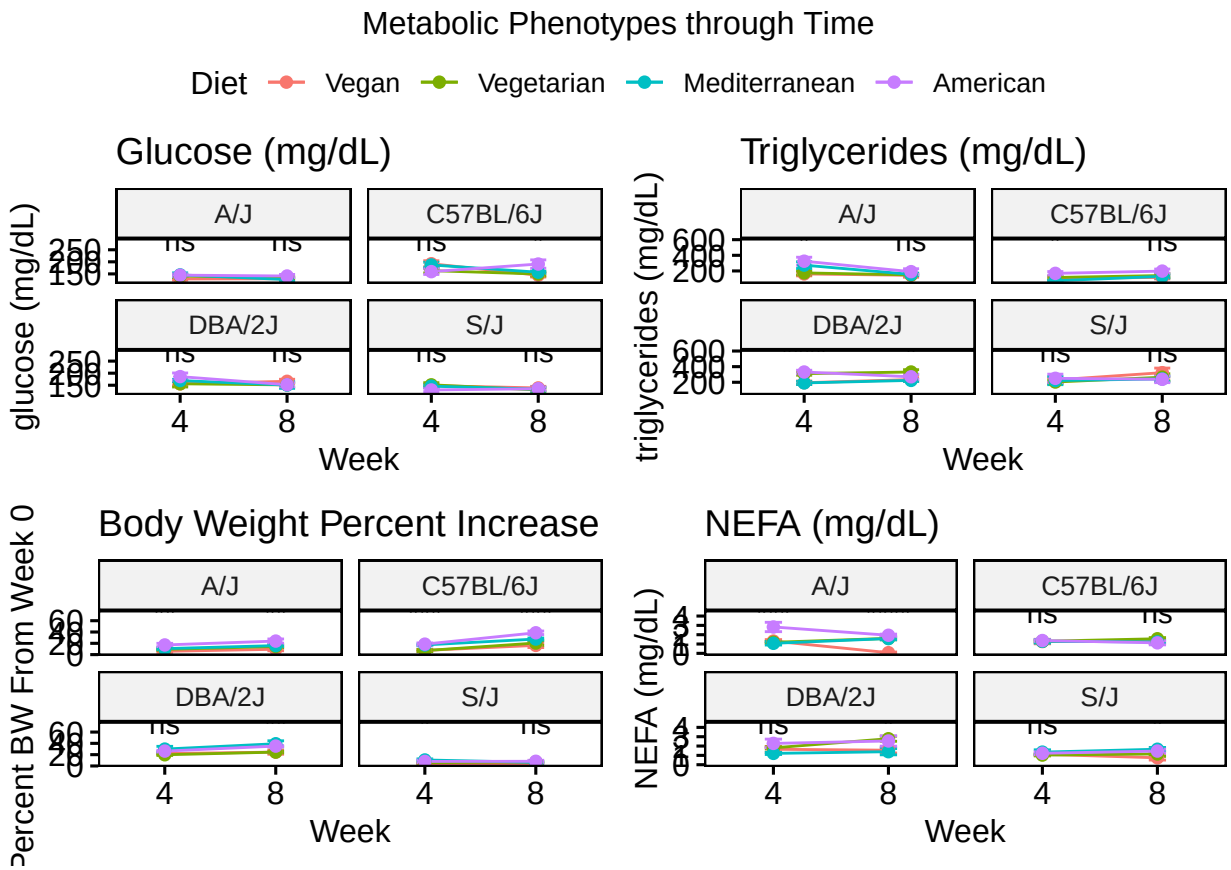
This a plot of one of contributing variables (BW_Mass or percent weight gained from Week 0) over time in the experiment. It is interesting to see the differential weight gains based on diet by strain. We see not only differential weight gains based on diet by strain, but also possibly bi-model weight gain outcomes by diet (see DBA/2J and C57BL6).

```
dat$Diet <- factor(dat$Diet, levels = c("Chow", "Vegan", "Vegetarian", "Mediterranean", "American"))
line <- ggline(dat, x = "Week", y = "PercBWincrease", color = "Diet", title = "Percent Body Weight Gain
line + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
```



We also looked at how our highest contributing variables changed over time. As Susanna pointed out, it is possible that some of those variables have a more important role at a different time point, suggesting possibly pathogensis roles that wouldn't otherwise be seen. We can see this most clearly in time point 4 for glucose for DBA/2J. But the are other examples here, and in the next slide with the other two highest contributing variables.

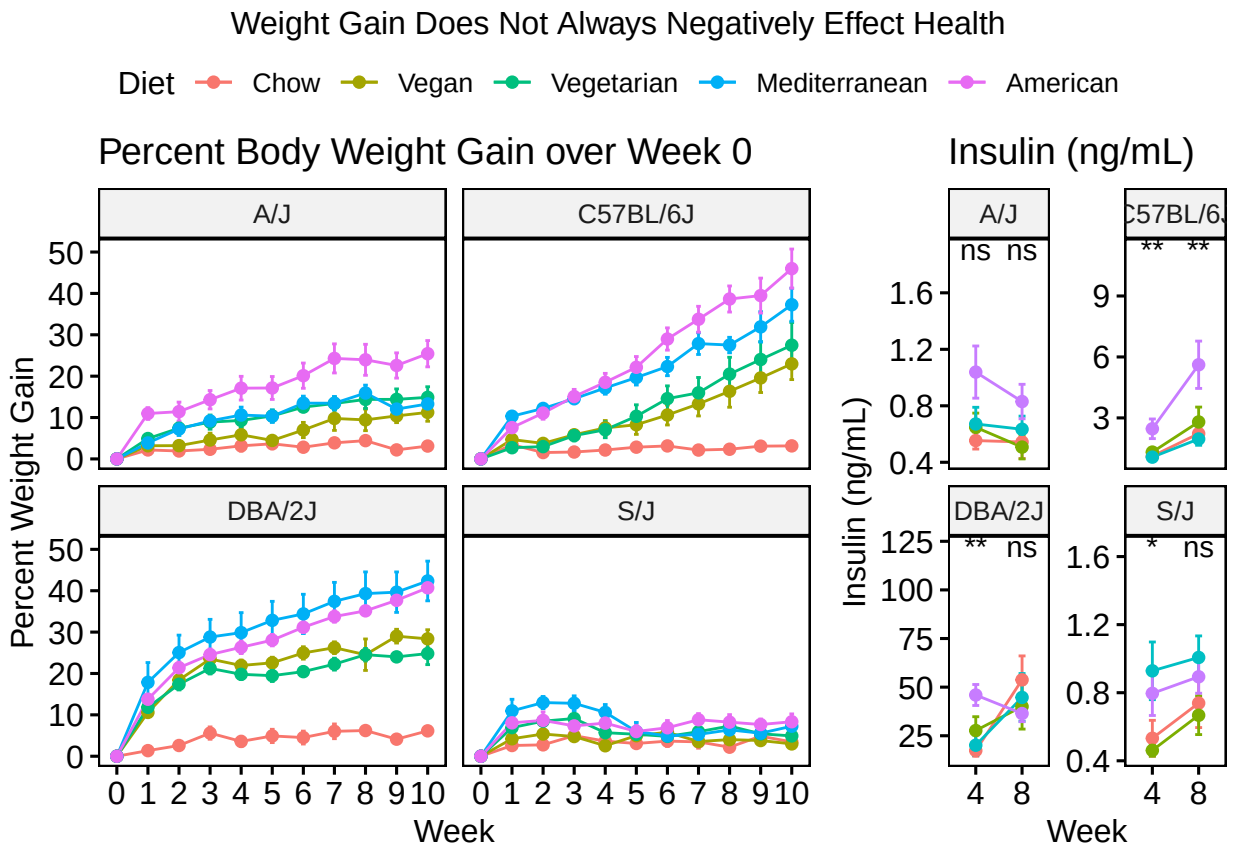
```
#line plot of top variables by time (Supplementary Figure 4)
PCA_dat$Diet <- factor(PCA_dat$Diet, levels = c("Chow", "Vegan", "Vegetarian", "Mediterranean", "America
glucose <- ggline(PCA_dat, x = "Week", y = "Glucose", color = "Diet", title = "Glucose (mg/dL)", xlab =
glucose <- glucose + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
trig <- ggline(PCA_dat, x = "Week", y = "Trig", color = "Diet", title = "Triglycerides (mg/dL)", xlab =
trig <- trig + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
percdiff <- ggline(PCA_dat, x = "Week", y = "PercBWincrease", color = "Diet", title = "Body Weight Perc
percdiff <- percdiff + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
insulin <- ggline(PCA_dat, x = "Week", y = "Insulin", color = "Diet", title = "Insulin (ng/mL)", xlab =
insulin <- insulin + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
nefa <- ggline(PCA_dat, x = "Week", y = "NEFA", color = "Diet", title = "NEFA (mg/dL)", xlab = "Week",
nefa <- nefa + stat_compare_means(aes(group = Diet), method = "anova", label = "p.signif")
topPCs <- ggarrange(glucose, trig, percdiff, nefa,ncol = 2, nrow =2, common.legend = TRUE)
annotate_figure(topPCs, top = text_grob("Metabolic Phenotypes through Time"))
```



Beginning to Equate These GXE Outcomes To Health Outcomes

To begin to relate these data to health outcomes, I've plotted body weight gain next to insulin through time. I did this on the suggestion that while weight gain and insulin differences definietely show GXE effects, they do not necessarily have poor health outcomes.

```
diffeffect <- ggarrange(line, insulin, widths = c(2, 1), common.legend = TRUE)
annotate_figure(diffeffect, top = text_grob("Weight Gain Does Not Always Negatively Effect Health"))
```



For example, you can see that DBA/2J gain weight pretty significantly over time but the "stereotypical" insulin resistence is not found. In fact, we see the opposite. There are other examples of these in the data if you take a look.

True Health Scores

So far, we've seen as in Barrington (2018), they developed a health outcome score based on Cohen's d of the difference in Betas based on American diet. Here (and there), we think that's a biased view of the GXE data as not all of our strains respond negatively to American diet. Indeed, looking at the one-way ANOVA results, we can see that not one diet is best over another given strain, so we're planning on creating a measure for health score that can maybe be referenced to the chow diet, or perhaps we can find some other way to create a distance measure between diets for a health score as the PCA clearly indicates GXE interactions that are not uniform across Strains and variables.