Practicals Day 1

CODE

Florence Mehl

February 26, 2025

- Nutrimouse dataset
- Unsupervised analysis
  - Genes dataset
    - Investigate distribution of data.
    - Perform PCA and investigate variances, sample distribution and variable relationship with plots.
    - Plot the explained variances
    - Observe the samples distributions in the space of the dimensions, what are the main sources of variation?
    - Which variables are responsible of the samples differences?
  - Lipids dataset
    - Perform PCA and investigate variances, sample distribution and variable relationship with plots.
    - Plot the explained variances
    - Observe the samples distributions in the space of the dimensions, what are the main sources of variation?
    - Which variables are responsible of the samples differences?
- Supervised analysis
  - PLS analysis between genes and lipids datasets
    - PLS canonical analysis
    - Samples distribution in the new reference (rotated axes) for each of the two blocks
    - Variables contribution in each data block to each dimension, after deflating more *important* variates
    - Compare samples distribution obtained from the regression and canonical PLS analyses.
  - PLS-DA analysis of the genes dataset to discriminate between genotypes
    - Plot the projection of samples on the latent variable
    - Which variables are responsible of the samples differences?
  - OPLS-DA analysis of the genes dataset to discriminate between genotypes
    - Plot the projection of samples on the latent variable
    - Which variables are responsible of the samples differences?

Nutrimouse dataset

The data sets come from a nutrigenomic study in the mouse (Martin et al., 2007) in which the effects of five regimens with contrasted fatty acid compositions on liver lipids and hepatic gene expression in mice were considered.

Two sets of variables were acquired on forty mice: - genes: expressions of 120 genes measured in liver cells, selected (among about 30,000) as potentially relevant in the context of the nutrition study. These expressions come from a nylon macroarray with radioactive labelling - lipids: concentrations (in percentages) of 21 hepatic fatty acids measured by gas chromatography

Biological units (mice) were cross-classified according to two factors experimental design (4 replicates): - genotype: 2-levels factor, wild-type (WT) and PPARalpha -/- (PPAR) - diet: 5-levels factor. Oils used for experimental diets preparation were corn and colza oils (50/50) for a reference diet (REF), hydrogenated coconut oil for a saturated fatty acid diet (COC), sunflower oil for an Omega6 fatty acid-rich diet (SUN), linseed oil for an Omega3-rich diet (LIN) and corn/colza/enriched fish oils for the FISH diet (43/43/14)

HIDE

```
data("nutrimouse")
genes <- nutrimouse$gene
lipids <- nutrimouse$lipid
metadata <- data.frame(genotype = nutrimouse$genotype, diet = nutrimouse$diet)
metadata$sample_name <- paste0(rownames(metadata), "_", metadata$genotype, "_", metadata$diet)
rownames(genes) <- metadata$sample_name
rownames(lipids) <- metadata$sample_name
```

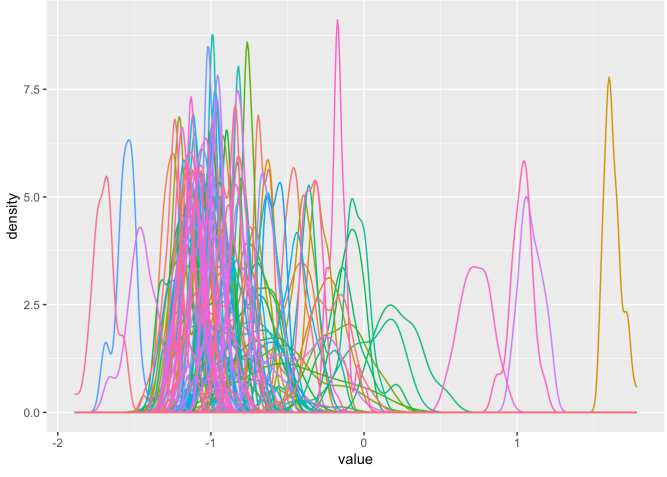
Unsupervised analysis

Genes dataset

Investigate distribution of data.

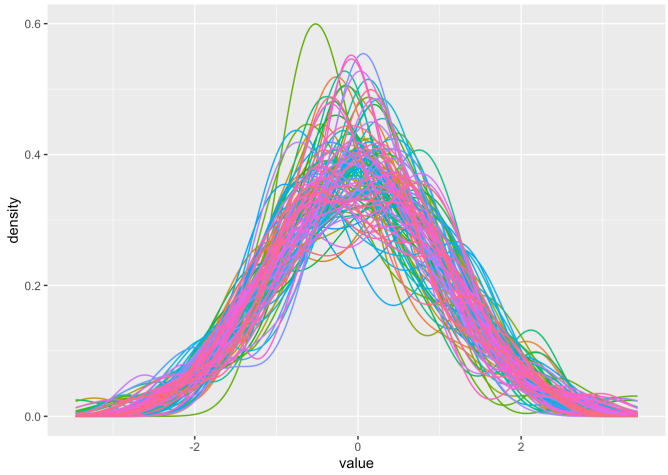
HIDE

```
genes.melt <- melt(nutrimouse$gene)

ggplot(genes.melt, aes(x=value, col=variable)) +
  geom_density() +
  guides(col=F)
```



HIDE

```
scaled.genes.melt <- melt(scale(nutrimouse$gene))
scaled.genes.melt <- scaled.genes.melt[,-1]
colnames(scaled.genes.melt) <- c("variable", "value")

ggplot(scaled.genes.melt, aes(x=value, col=variable)) +
  geom_density() +
  guides(col=F)
```

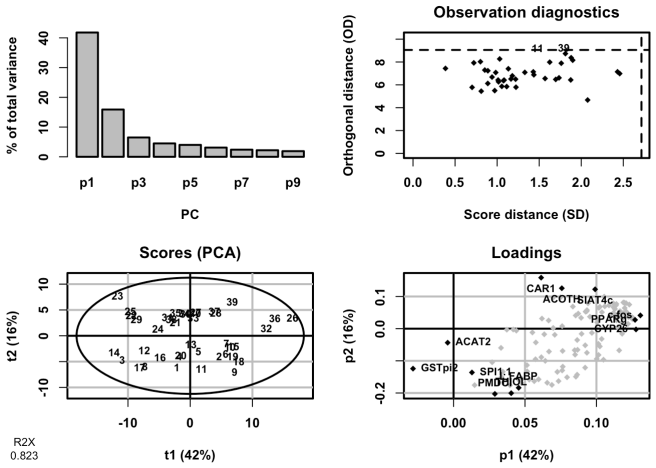


Perform PCA and investigate variances, sample distribution and variable relationship with plots.

HIDE

```
# run analysis
PCA_genes_res <- opls(x = nutrimouse$gene,
                      predI = 9,
                      crossvalI = 7,
                      permI = 100)
```

```
## PCA
## 40 samples x 120 variables
## standard scaling of predictors
##       R2X(cum) pre ort
## Total    0.823   9   0
```

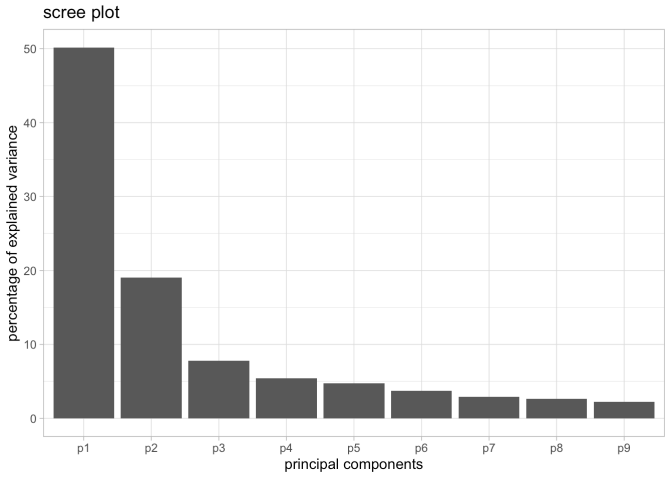


Plot the explained variances

- scree plot

HIDE

```
# saliences

variances_genes <- data.frame(variance = PCA_genes_res@pcaVarVn)
variances_genes$Dim <- rownames(variances_genes)

ggplot(variances_genes, aes(x=Dim, y=variance)) +
  geom_bar(stat = "identity") +
  theme_light() +
  labs(x = "principal components",
       y = "percentage of explained variance",
       title = "scree plot")
```

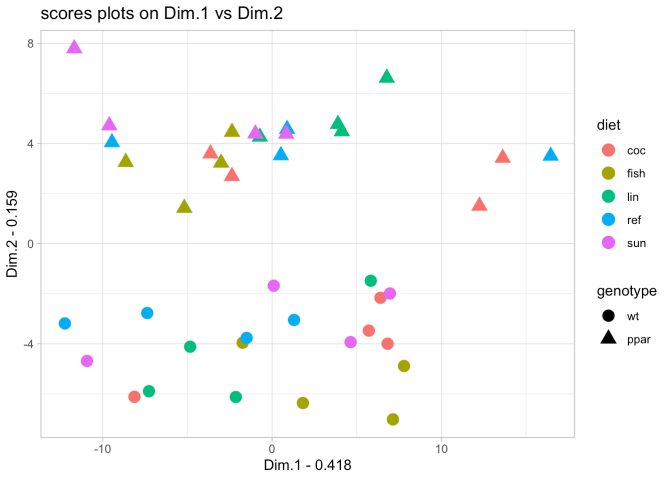


Observe the samples distributions in the space of the dimensions, what are the main sources of variation?

- plot scores on Dim.1 vs Dim.2 with explained variance on axes
- plot scores on Dim.3 vs Dim.4 with explained variance on axes

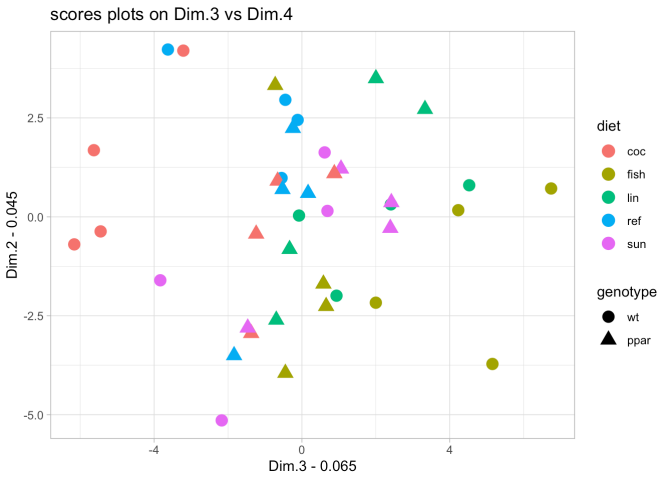
HIDE

```
scores_genes <- data.frame(metadata, PCA_genes_res@scoreMN)

ggplot(scores_genes, aes(x=p1, y=p2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(x=paste0("Dim.1 - ", PCA_genes_res@modelDF$R2X[1]),
       y=paste0("Dim.2 - ", PCA_genes_res@modelDF$R2X[2]),
       title = "scores plots on Dim.1 vs Dim.2") +
  theme_light()
```



HIDE

```
ggplot(scores_genes, aes(x=p3, y=p4, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(x=paste0("Dim.3 - ", PCA_genes_res@modelDF$R2X[3]),
       y=paste0("Dim.2 - ", PCA_genes_res@modelDF$R2X[4]),
       title = "scores plots on Dim.3 vs Dim.4") +
  theme_light()
```

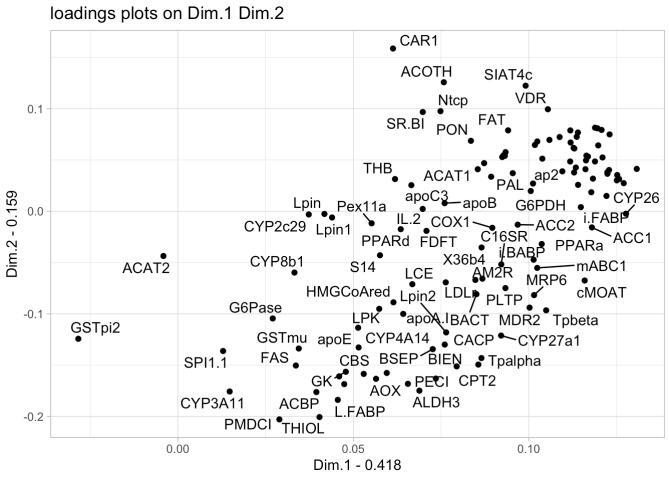


Which variables are responsible of the samples differences?

- plot loadings on Dim.1 vs Dim.2 with explained variance on axes
- plot loadings on Dim.3 vs Dim.4 with explained variance on axes

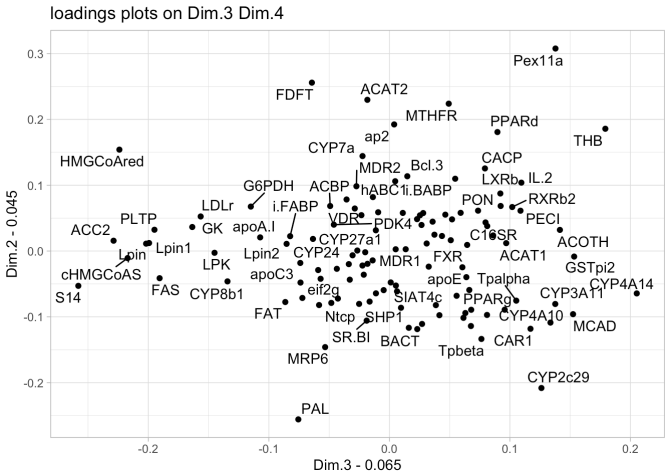
HIDE

```
loadings_genes <- data.frame(PCA_genes_res@loadingMN)
loadings_genes$variable <- rownames(loadings_genes)

ggplot(loadings_genes, aes(x=p1, y=p2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(x=paste0("Dim.1 - ", PCA_genes_res@modelDF$R2X[1]),
       y=paste0("Dim.2 - ", PCA_genes_res@modelDF$R2X[2]),
       title = "loadings plots on Dim.1 Dim.2") +
  theme_light()
```



HIDE

```
ggplot(loadings_genes, aes(x=p3, y=p4, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(x=paste0("Dim.3 - ", PCA_genes_res@modelDF$R2X[3]),
       y=paste0("Dim.2 - ", PCA_genes_res@modelDF$R2X[4]),
       title = "loadings plots on Dim.3 Dim.4") +
  theme_light()
```



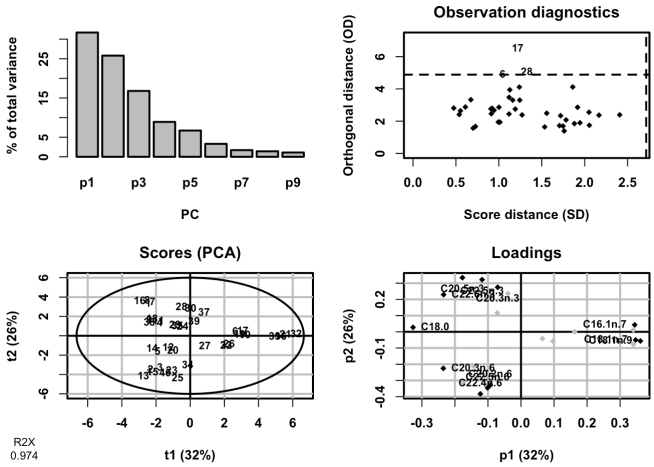
Lipids dataset

Perform PCA and investigate variances, sample distribution and variable relationship with plots.

HIDE

```
# run analysis
PCA_lipids_res <- opls(x = nutrimouse$lipid,
                       predI = 9,
                       crossvalI = 7,
                       permI = 100)
```

```
## PCA
## 40 samples x 21 variables
## standard scaling of predictors
##       R2X(cum) pre ort
## Total    0.974   9   0
```

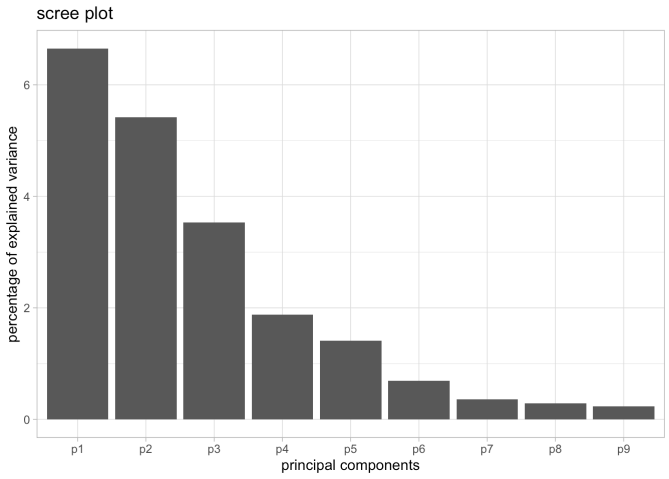


Plot the explained variances

- scree plot

HIDE

```
# saliences

variances_lipids <- data.frame(variance = PCA_lipids_res@pcaVarVn)
variances_lipids$Dim <- rownames(variances_lipids)

ggplot(variances_lipids, aes(x=Dim, y=variance)) +
  geom_bar(stat = "identity") +
  theme_light() +
  labs(x = "principal components",
       y = "percentage of explained variance",
       title = "scree plot")
```

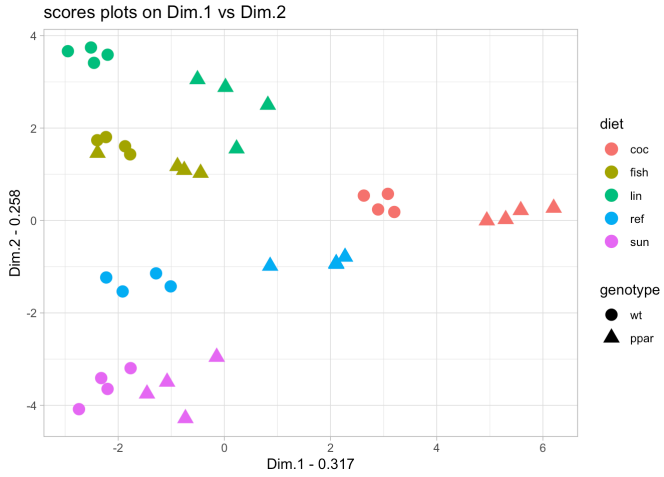


Observe the samples distributions in the space of the dimensions, what are the main sources of variation?

- plot scores on Dim.1 vs Dim.2 with explained variance on axes
- plot scores on Dim.3 vs Dim.4 with explained variance on axes

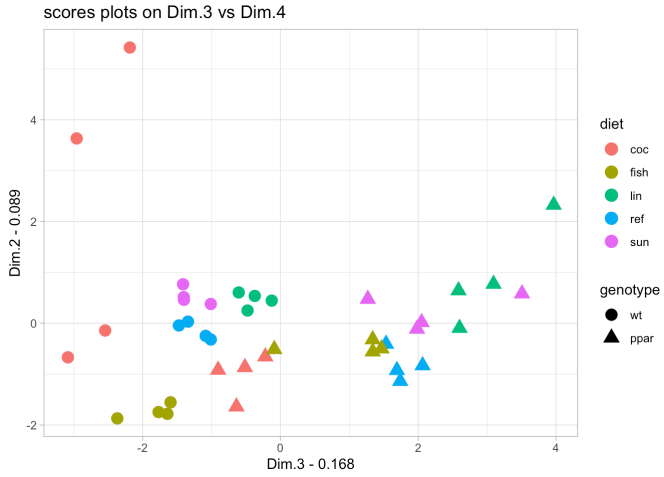
HIDE

```
scores_lipids <- data.frame(metadata, PCA_lipids_res@scoreMN)

ggplot(scores_lipids, aes(x=p1, y=p2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(x=paste0("Dim.1 - ", PCA_lipids_res@modelDF$R2X[1]),
       y=paste0("Dim.2 - ", PCA_lipids_res@modelDF$R2X[2]),
       title = "scores plots on Dim.1 vs Dim.2") +
  theme_light()
```



HIDE

```
ggplot(scores_lipids, aes(x=p3, y=p4, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(x=paste0("Dim.3 - ", PCA_lipids_res@modelDF$R2X[3]),
       y=paste0("Dim.2 - ", PCA_lipids_res@modelDF$R2X[4]),
       title = "scores plots on Dim.3 vs Dim.4") +
  theme_light()
```

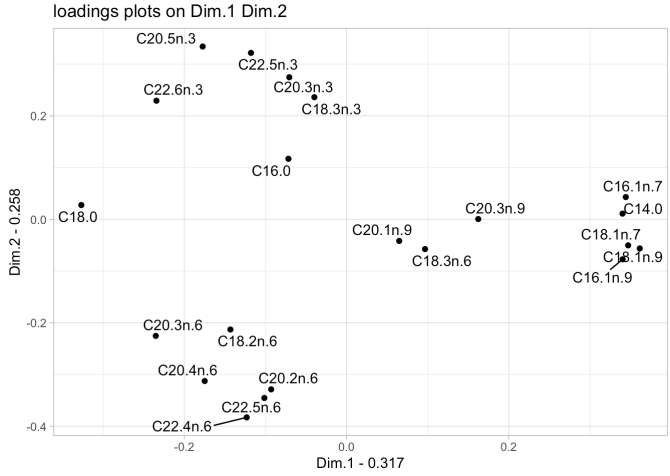


Which variables are responsible of the samples differences?

- plot loadings on Dim.1 vs Dim.2 with explained variance on axes
- plot loadings on Dim.3 vs Dim.4 with explained variance on axes

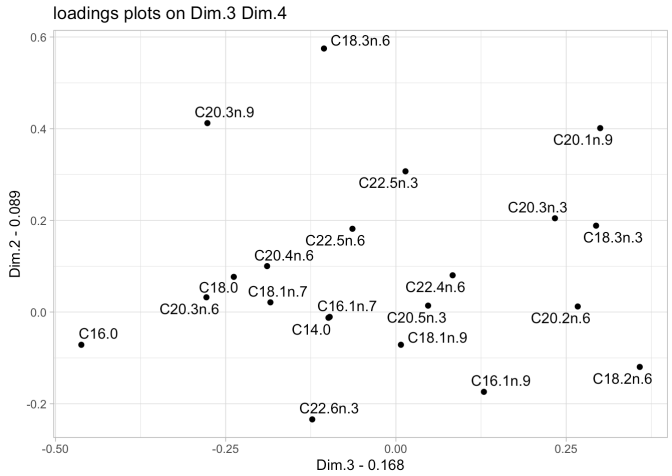
HIDE

```
loadings_lipids <- data.frame(PCA_lipids_res@loadingMN)
loadings_lipids$variable <- rownames(loadings_lipids)

ggplot(loadings_lipids, aes(x=p1, y=p2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(x=paste0("Dim.1 - ", PCA_lipids_res@modelDF$R2X[1]),
       y=paste0("Dim.2 - ", PCA_lipids_res@modelDF$R2X[2]),
       title = "loadings plots on Dim.1 Dim.2") +
  theme_light()
```



HIDE

```
ggplot(loadings_lipids, aes(x=p3, y=p4, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(x=paste0("Dim.3 - ", PCA_lipids_res@modelDF$R2X[3]),
       y=paste0("Dim.2 - ", PCA_lipids_res@modelDF$R2X[4]),
       title = "loadings plots on Dim.3 Dim.4") +
  theme_light()
```



Supervised analysis

PLS analysis between genes and lipids datasets

PLS canonical analysis

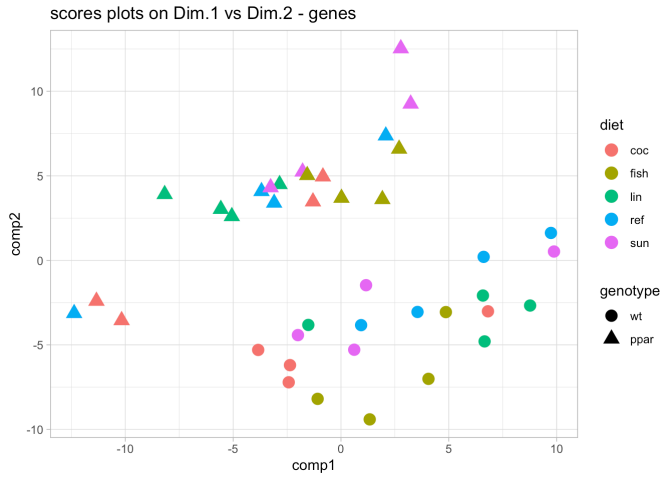
HIDE

```
library(mixOmics)

# run analysis
PLS_cano_res <- pls(X=nutrimouse$gene, Y=nutrimouse$lipid, ncomp=2, scale=TRUE, mode="canonical")
```

Samples distribution in the new reference (rotated axes) for each of the two blocks

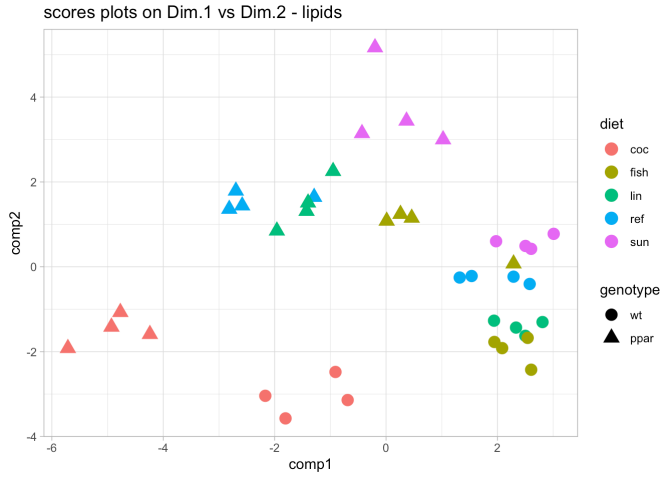
HIDE

```
PLS_cano_scores_genes <- data.frame(metadata, PLS_cano_res$variates$X)

ggplot(PLS_cano_scores_genes, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - genes") +
  theme_light()
```



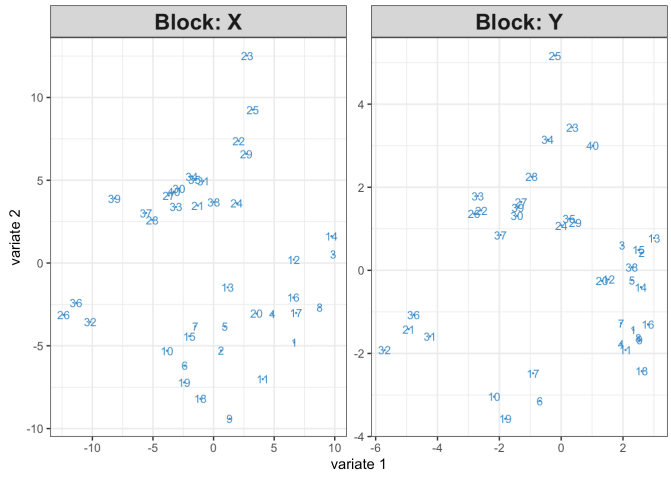
HIDE

```
PLS_cano_scores_lipids <- data.frame(metadata, PLS_cano_res$variates$Y)

ggplot(PLS_cano_scores_lipids, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - lipids") +
  theme_light()
```



HIDE

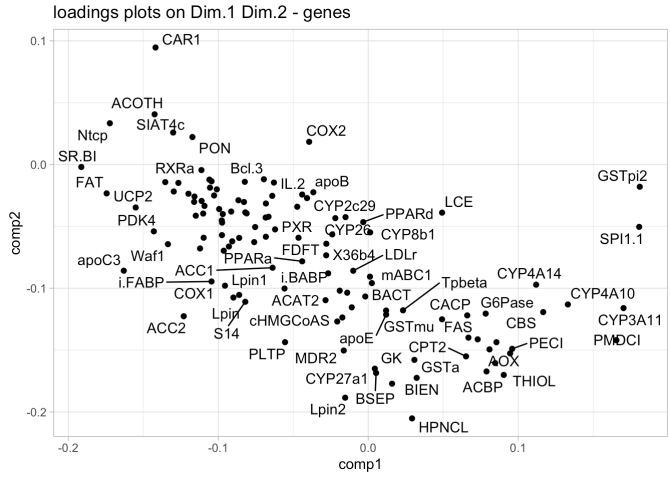
```
# or with function from mixomics
plotIndiv(PLS_cano_res)
```



Variables contribution in each data block to each dimension, after deflating more *important* variates

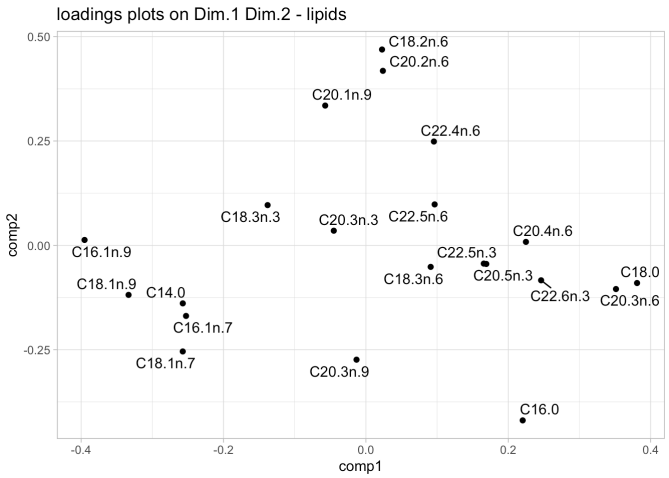
HIDE

```
PLS_cano_loadings_genes <- data.frame(PLS_cano_res$loadings$X)
PLS_cano_loadings_genes$variable <- rownames(PLS_cano_loadings_genes)

ggplot(PLS_cano_loadings_genes, aes(x=comp1, y=comp2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - genes") +
  theme_light()
```



HIDE

```
PLS_cano_loadings_lipids <- data.frame(PLS_cano_res$loadings$Y)
PLS_cano_loadings_lipids$variable <- rownames(PLS_cano_loadings_lipids)

ggplot(PLS_cano_loadings_lipids, aes(x=comp1, y=comp2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - lipids") +
  theme_light()
```



Observe the difference between the two modes *regression* and *canonical* of PLS.

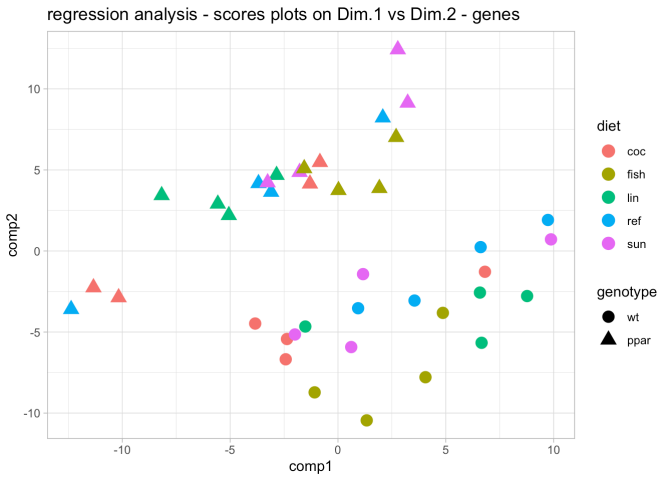
HIDE

```
PLS_reg_res <- pls(X=nutrimouse$gene, Y=nutrimouse$lipid, ncomp=2, scale=TRUE, mode="regression")
```

Compare samples distribution obtained from the regression and canonical PLS analyses.

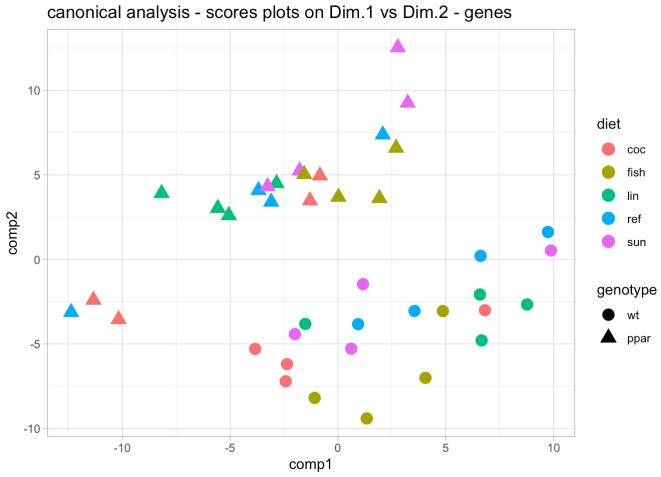
HIDE

```
PLS_reg_scores_genes <- data.frame(metadata, PLS_reg_res$variates$X)

ggplot(PLS_reg_scores_genes, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "regression analysis - scores plots on Dim.1 vs Dim.2 - genes") +
  theme_light()
```



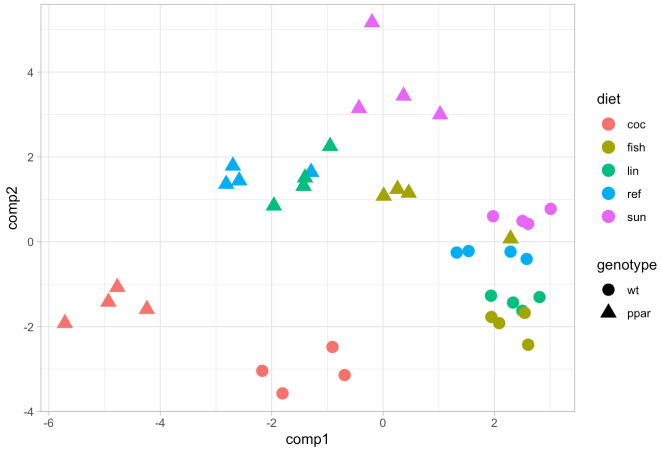
HIDE

```
ggplot(PLS_cano_scores_genes, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "canonical analysis - scores plots on Dim.1 vs Dim.2 - genes") +
  theme_light()
```



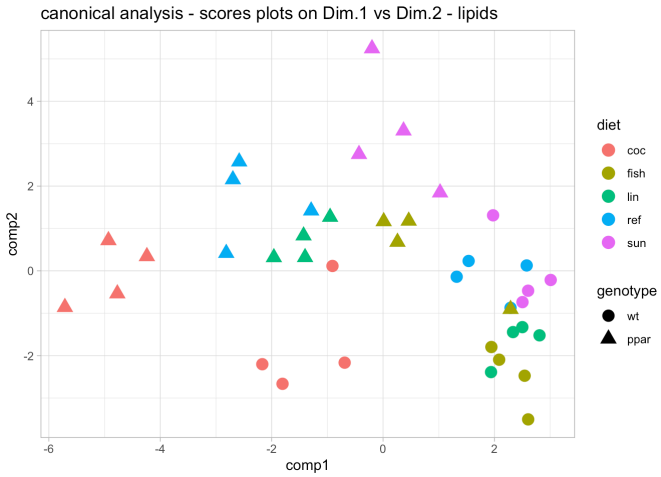
HIDE

```
PLS_reg_scores_lipids <- data.frame(metadata, PLS_reg_res$variates$Y)

ggplot(PLS_cano_scores_lipids, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "regression analysis - scores plots on Dim.1 vs Dim.2 - lipids") +
  theme_light()
```



HIDE

```
ggplot(PLS_reg_scores_lipids, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "canonical analysis - scores plots on Dim.1 vs Dim.2 - lipids") +
  theme_light()
```

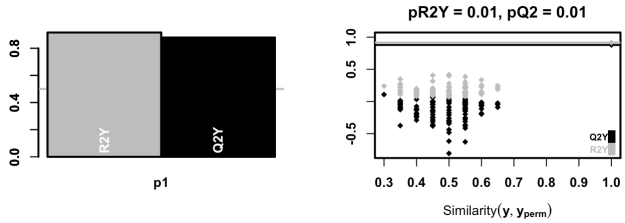


PLS-DA analysis of the genes dataset to discriminate between genotypes

HIDE

```
PLSDA_genes_genotype <- opls(x = nutrimouse$gene,
                             y = metadata$genotype,
                             predI = NA,
                             permI = 100)
```

```
## PLS-DA
## 40 samples x 120 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total    0.158    0.916   0.879 0.149   1   0 0.01 0.01
```

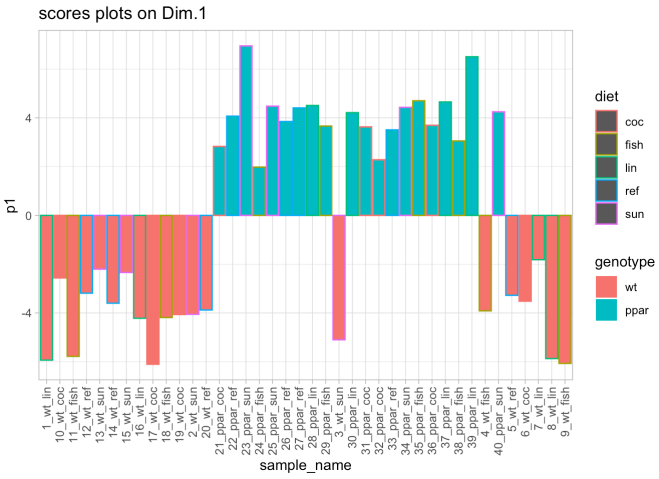


Plot the projection of samples on the latent variable

- plot scores on Dim.1

HIDE

```
scores_genes_genotype <- data.frame(metadata, PLSDA_genes_genotype@scoreMN)

ggplot(scores_genes_genotype, aes(x=sample_name, y=p1, fill=genotype, col = diet)) +
  geom_bar(stat = "identity") +
  labs(title = "scores plots on Dim.1") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

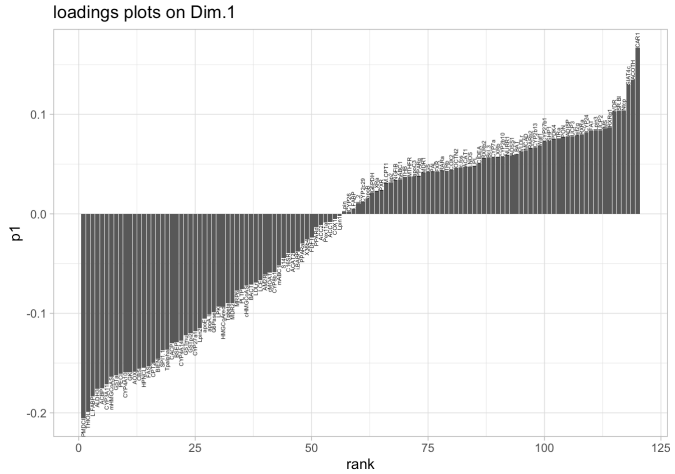


Which variables are responsible of the samples differences?

- plot loadings on Dim.1

HIDE

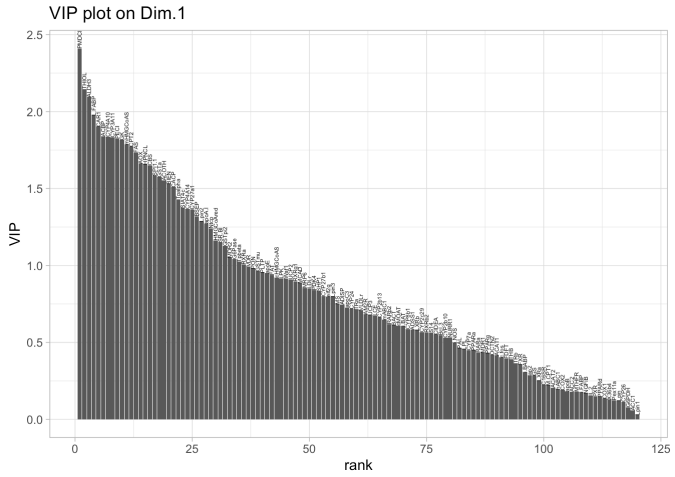
```
loadings_genes_genotype <- data.frame(PLSDA_genes_genotype@loadingMN)
loadings_genes_genotype$variable <- rownames(loadings_genes_genotype)
loadings_genes_genotype <- loadings_genes_genotype[order(loadings_genes_genotype$p1),]
loadings_genes_genotype$rank <- seq(1,nrow(loadings_genes_genotype))

ggplot(loadings_genes_genotype, aes(x=rank, y=p1, label=variable)) +
  geom_bar(stat = "identity") +
  geom_text(angle=90, size=1.5, hjust=ifelse(loadings_genes_genotype$p1 <0, 1,0)) +
  labs(title = "loadings plots on Dim.1") +
  theme_light()
```



- plot VIP on Dim.1

HIDE

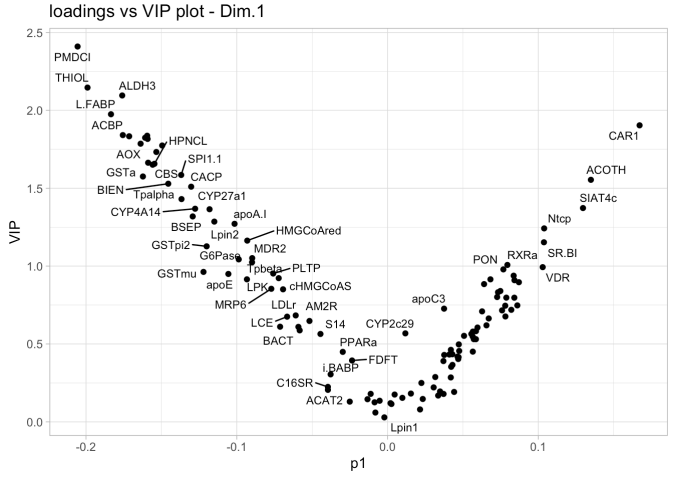
```
VIP_genes_genotype <- data.frame(VIP = PLSDA_genes_genotype@vipVn)
VIP_genes_genotype$variable <- rownames(VIP_genes_genotype)
VIP_genes_genotype <- VIP_genes_genotype[order(VIP_genes_genotype$VIP, decreasing = T),]
VIP_genes_genotype$rank <- seq(1,nrow(loadings_genes_genotype))

ggplot(VIP_genes_genotype, aes(x=rank, y=VIP, label=variable)) +
  geom_bar(stat = "identity") +
  geom_text(angle=90, size=1.5, hjust=0) +
  labs(title = "VIP plot on Dim.1") +
  theme_light()
```



- plot loadings vs VIP on Dim.1

HIDE

```
loadings_VIP_genes_genotype <- merge(loadings_genes_genotype[,-3], VIP_genes_genotype[,-3])

ggplot(loadings_VIP_genes_genotype, aes(x=p1, y=VIP, label=variable)) +
  geom_point() +
  geom_text_repel(size=3) +
  labs(title = "loadings vs VIP plot - Dim.1") +
  theme_light()
```

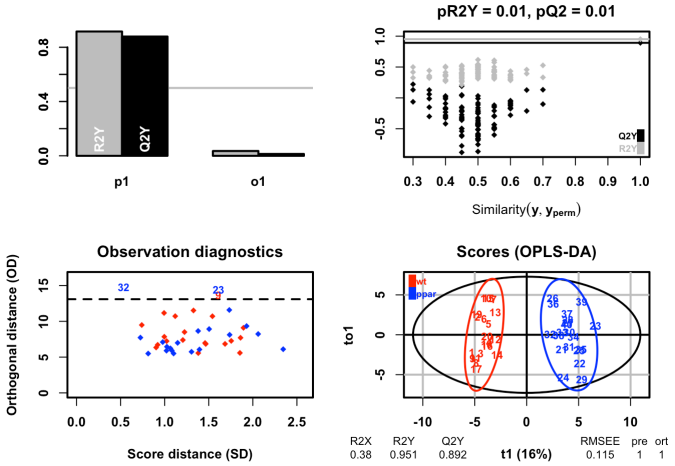


OPLS-DA analysis of the genes dataset to discriminate between genotypes

HIDE

```
OPLSDA_genes_genotype <- opls(x = nutrimouse$gene,
                              y = metadata$genotype,
                              predI = 1,
                              orthoI = 1,
                              permI = 100)
```

```
## OPLS-DA
## 40 samples x 120 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total     0.38    0.951   0.892 0.115   1   1 0.01 0.01
```

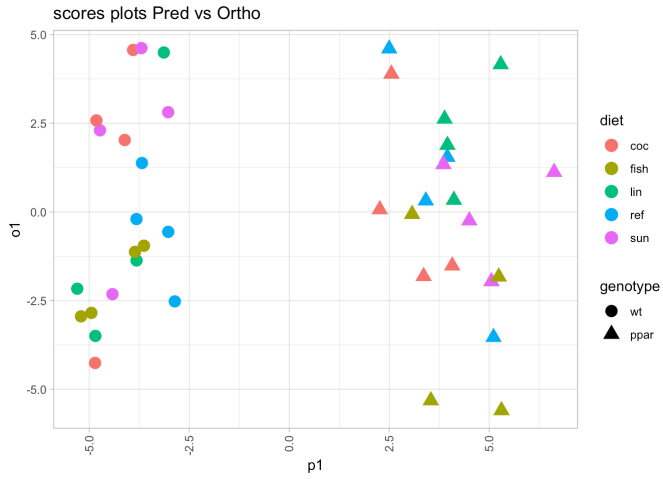


Plot the projection of samples on the latent variable

- plot scores on Pred vs Ortho

HIDE

```
oplsda_scores_genes_genotype <- data.frame(metadata, p1= OPLSDA_genes_genotype@scoreMN, o1=OPLSDA_genes_genotype@orthoScoreMN)

ggplot(oplsda_scores_genes_genotype, aes(x=p1, y=o1, shape=genotype, col = diet)) +
  geom_point(size=4) +
  labs(title = "scores plots Pred vs Ortho") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

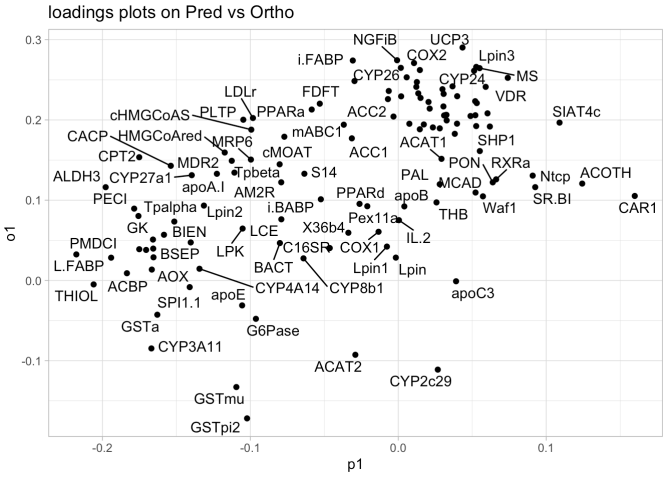


Which variables are responsible of the samples differences?

- plot loadings on Pred vs Ortho

HIDE

```
oplsda_loadings_genes_genotype <- data.frame(p1=OPLSDA_genes_genotype@loadingMN, o1=OPLSDA_genes_genotype@orthoLoadingMN)
oplsda_loadings_genes_genotype$variable <- rownames(oplsda_loadings_genes_genotype)

ggplot(oplsda_loadings_genes_genotype, aes(x=p1, y=o1, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Pred vs Ortho") +
  theme_light()
```



- plot loadings vs VIP on Pred

HIDE

```
oplsda_loadings_VIP_genes_genotype <- data.frame(oplsda_loadings_genes_genotype, VIP = OPLSDA_genes_genotype@vipVn)

ggplot(oplsda_loadings_VIP_genes_genotype, aes(x=p1, y=VIP, label=variable)) +
  geom_point() +
  geom_text_repel(size=2, max.overlaps = 25, segment.size=.2) +
  labs(title = "loadings vs VIP plot - Dim.1") +
  theme_light()
```