5. PCA genes - Lab

Fay

2022-10-08

Always change the knitting directory to the working directory! # Load libraries

```
library(tidyverse)
library(dplyr)
library(stringr)
library(FactoMineR)
library(reshape2)
library(corrplot)
library(factoextra)
library(lmtest)
library(ggpubr)
library(janitor)
library(pheatmap)
library(visdat)
```

Load data

I am using a normalized and imputed data set.

```
hm <- read.csv("output_data/2.imputed_MICE_data_set.csv")
```

vectors for selecting

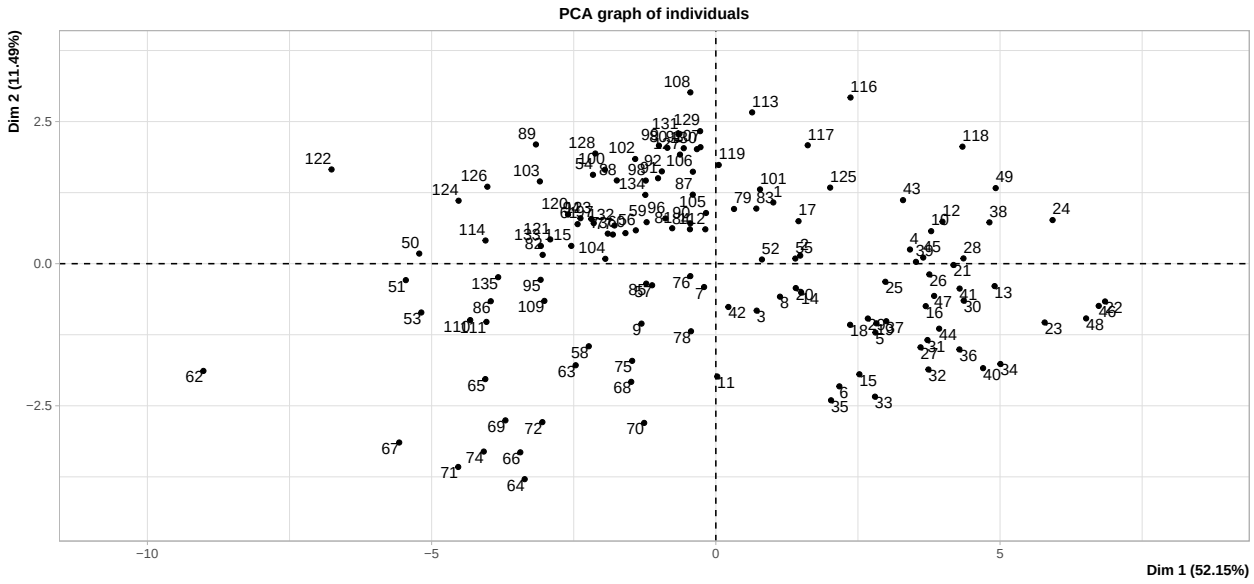
```
Gene_lab   <- c("IFNy", "CXCR3", "IL.6", "IL.13",
                "IL1RN","CASP1", "CXCL9", "IDO1", "IRGM1", "MPO",
                "MUC2", "MUC5AC", "MYD88", "NCR1", "PRF1", "RETNLB", "SOCS1",
                "TICAM1", "TNF") # "IL.12", "IRG6")
```

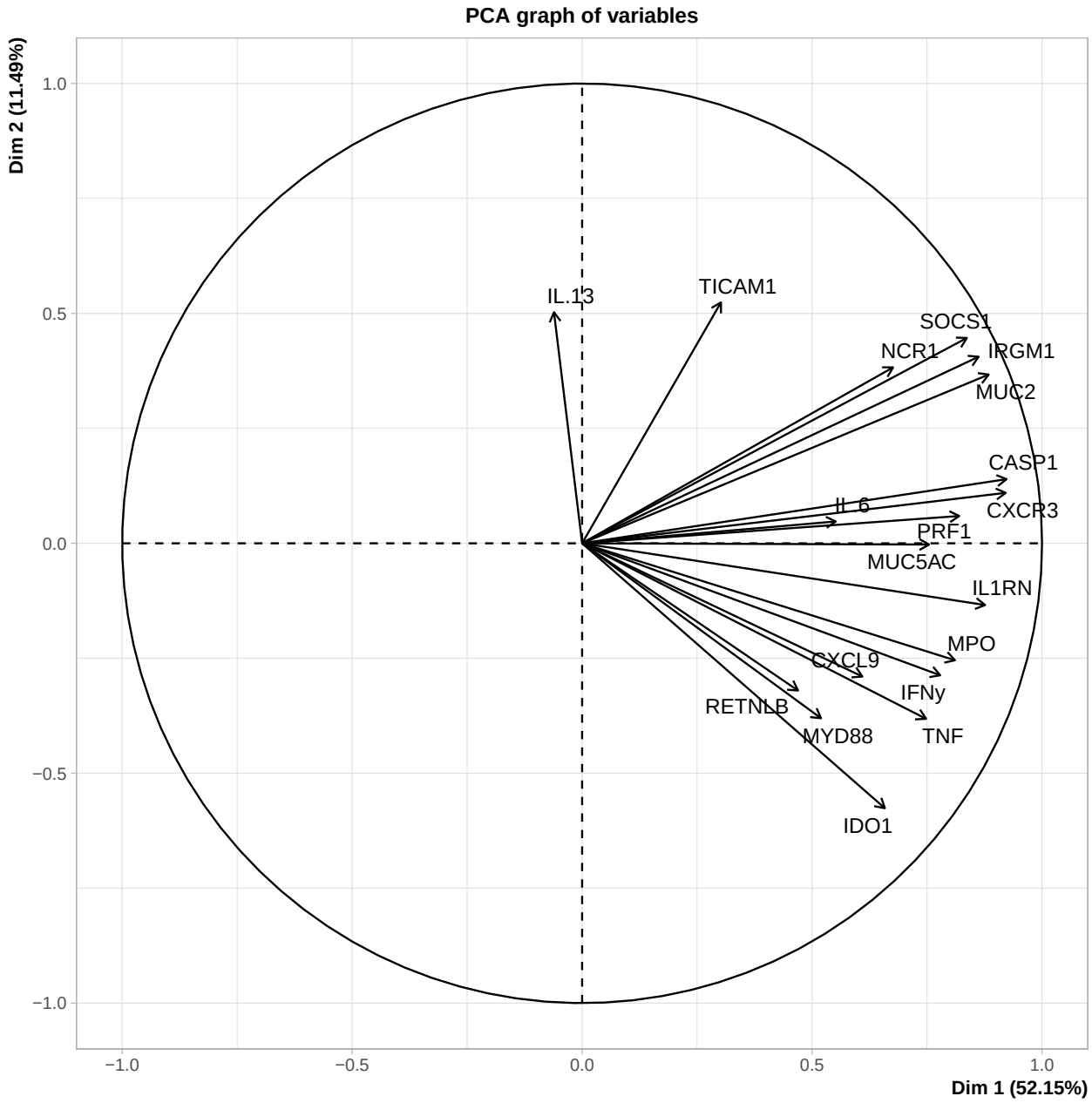
PCA on the lab genes

Firts I am preparing and cleaning the data before initiating the PCA.

```
#select the genes and lab muce
lab <- hm %>%
  dplyr::filter(origin == "Lab", Position == "mLN") #selecting for mln to avoid
# duplicates

lab <- unique(lab)

gene <- lab %>%
  dplyr::select(c(Mouse_ID, all_of(Gene_lab)))
```

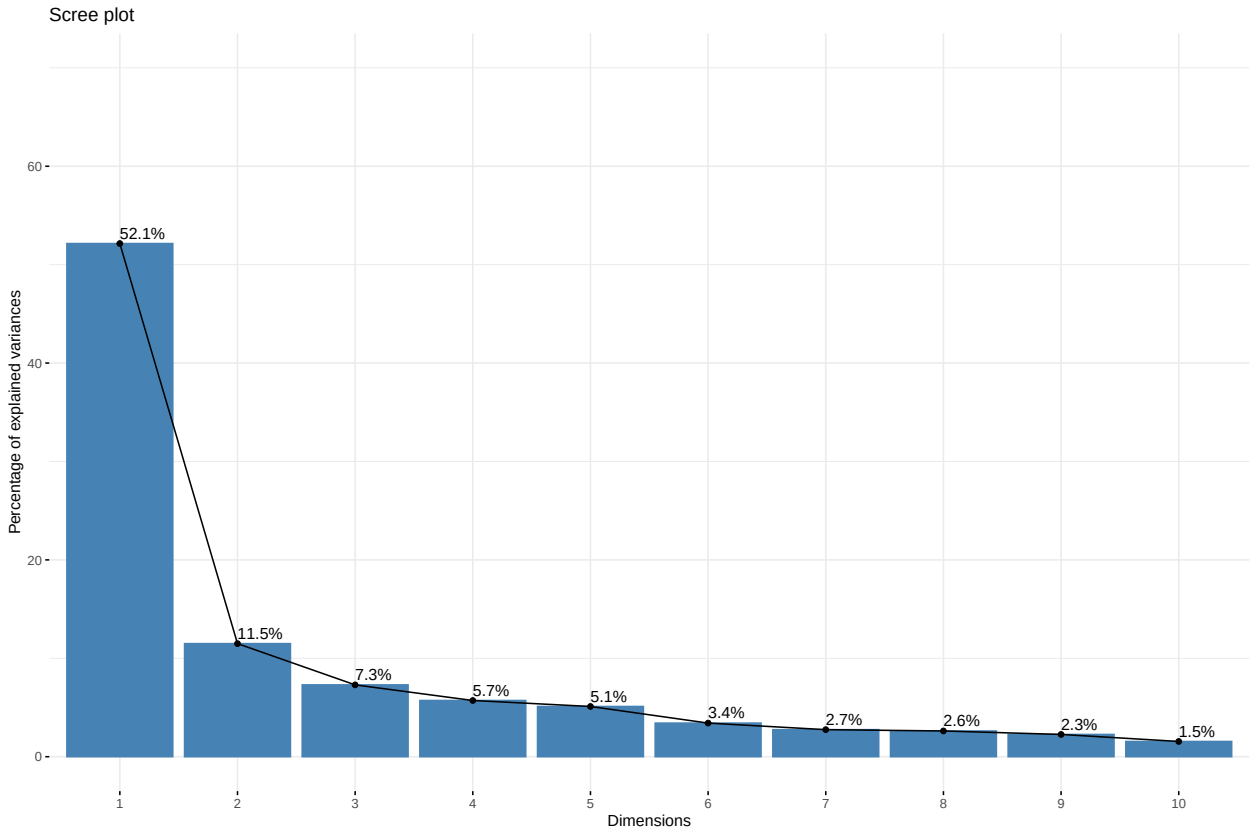
```
genes <- unique(gene)

genes <- genes[, -1]

#remove rows with only nas
genes <- genes[,colSums(is.na(genes))<nrow(genes)]

#remove colums with only nas
genes <- genes[rowSums(is.na(genes)) != ncol(genes), ]

#select same rows in the first table
gene <- gene[row.names(genes), ]

# we can now run a normal pca on the complete data set
res.pca <- PCA(genes)
```





How much does each dimension contribute to variabce?

```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 70))
```

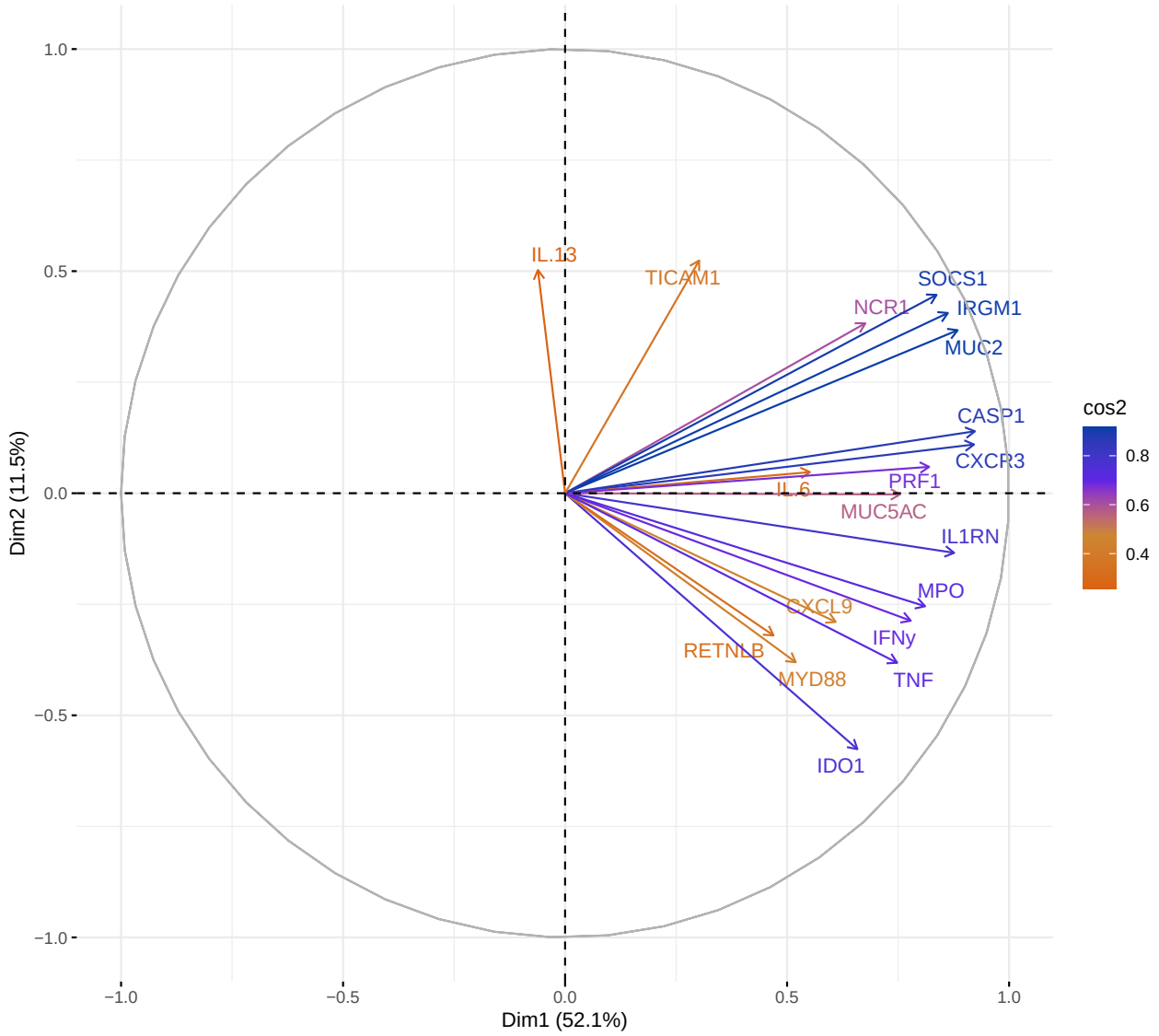


"Principal Component Analysis of Immune Gene Expression in Laboratory-Controlled Eimeria Infections"

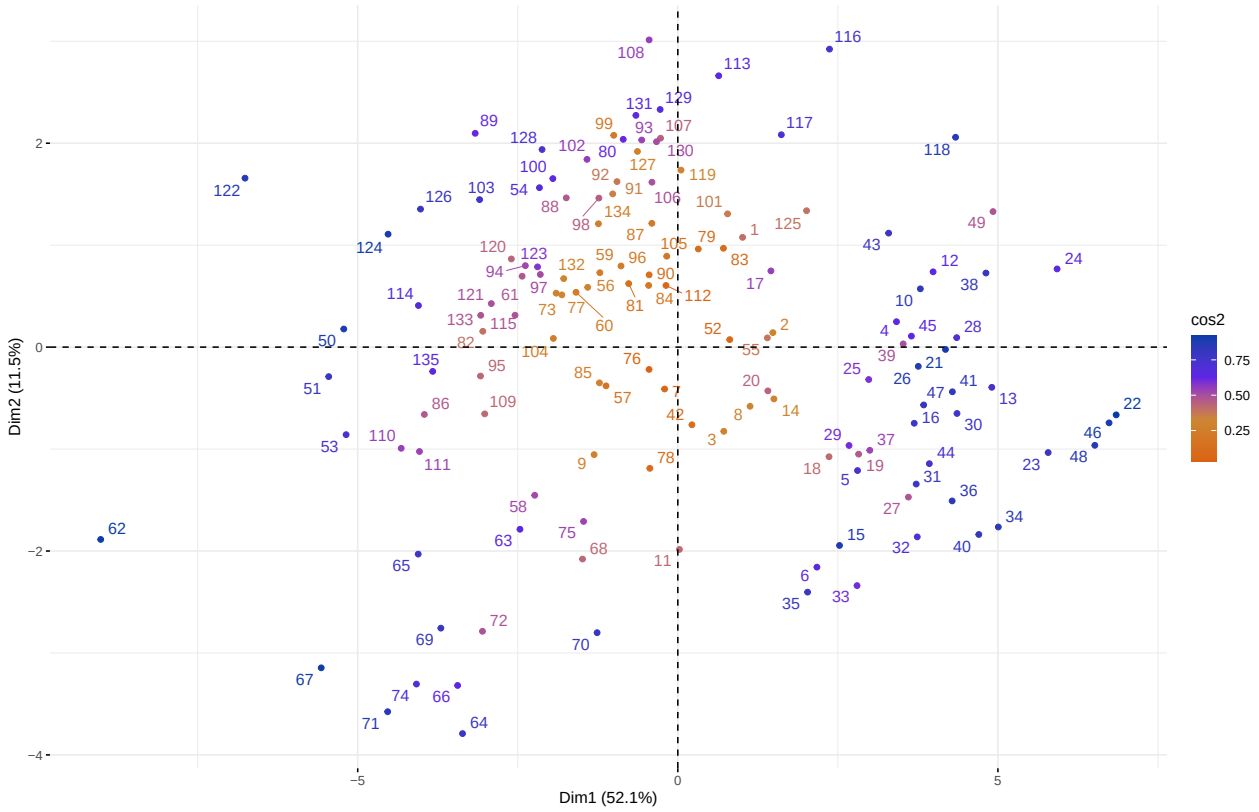
Part of the first figure in the publication in colour blind friendly colours.

Description: In this Principal Component Analysis plot, each point represents an individual mouse's immune response to a controlled Eimeria spp. infection, as determined by the expression levels of 19 key immune genes. The horizontal (PC1) and vertical (PC2) axes represent the two principal components that account for the largest possible variance in the gene expression data. In essence, these axes reveal the primary patterns of variation in immune gene expression across our samples.

```
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#DB6212", "#CC8733", "#5f25e6", "#073DA8"),
             repel = TRUE, title = "")
```

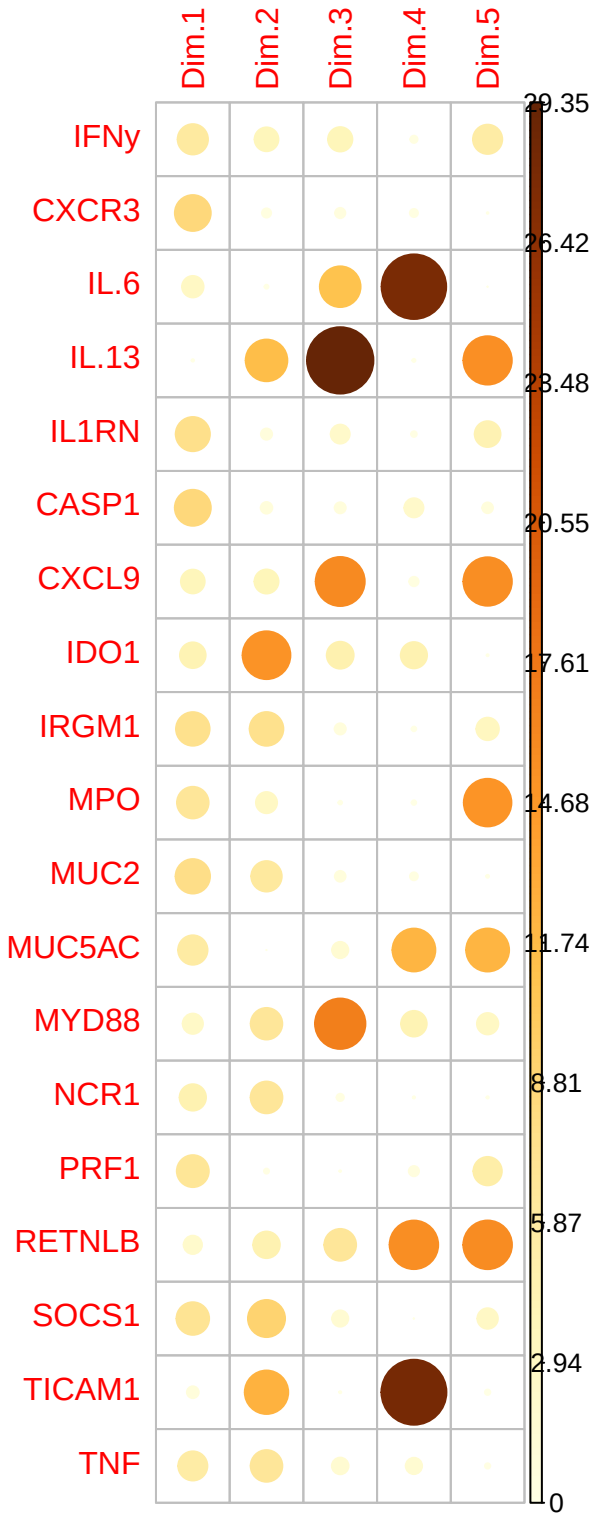



```
fviz_pca_ind(res.pca, col.ind = "cos2",
                  gradient.cols = c("#DB6212", "#CC8733", "#5f25e6", "#073DA8"),
                  repel = TRUE, title = "")
```



Adding the PC Eigenvectors to the data set.

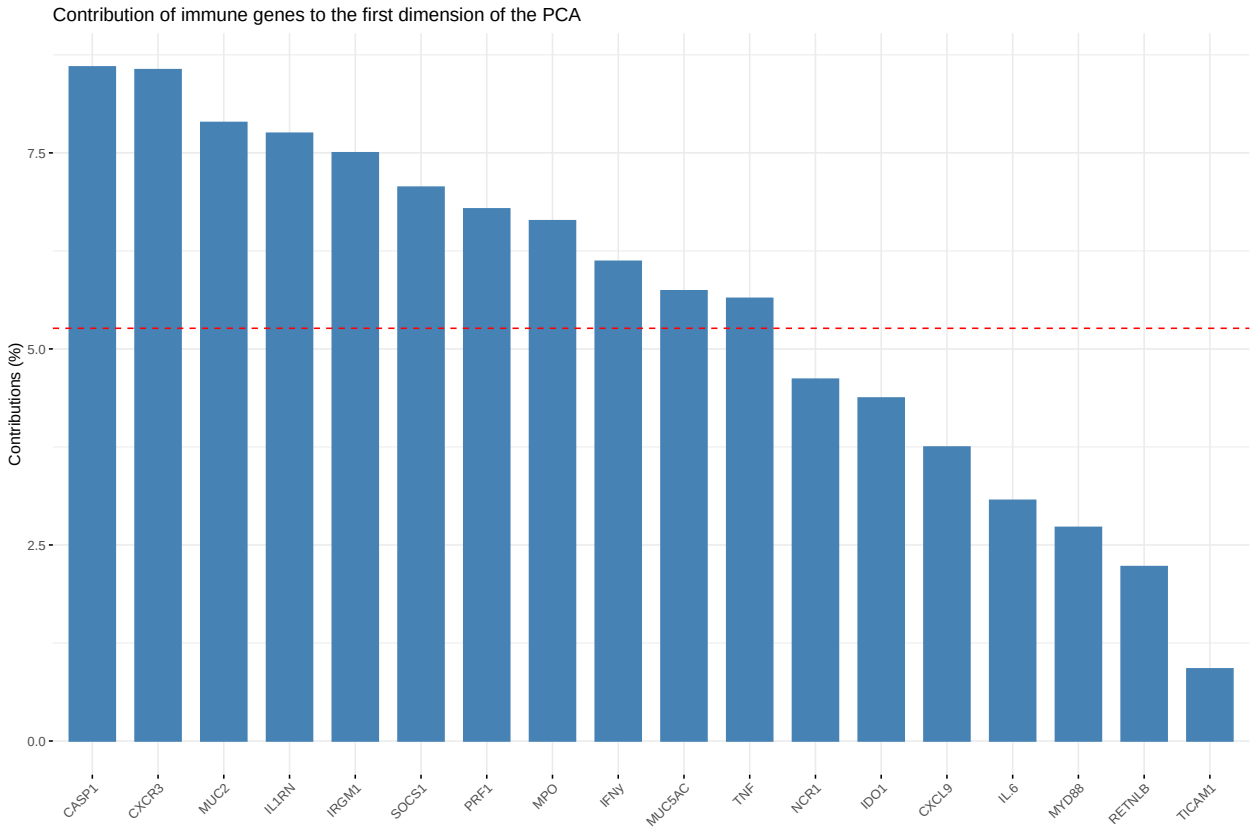
```
#It's possible to use the function corrplot() [corrplot package] to highlight
#the most contributing variables for each dimension:
var.contrib <- res.pca$var$contrib
corrplot(var.contrib, is.corr=FALSE)
```



The function fviz_contrib() [factoextra package] can be used to draw a bar plot of variable contributions. If your data contains many variables, you can decide to show only the top contributing variables. The R code below shows the top 10 variables contributing to the principal components:

Contributions to the first dimension

```
# Contributions of variables to PC1
fviz_contrib(res.pca, choice = "var", axes = 1, top = 18,
             title = "Contribution of immune genes to the first dimension of the PCA")
```




```
# res.pca$var$contrib
```

Title: "Contribution of Immune Genes to the First Principal Component of Gene Expression Analysis"

Description:

"The plot represents the contribution of each of the 19 immune genes to the first principal component (PC1) of a principal component analysis based on gene expression data derived from controlled infection experiments in laboratory mice.

Among these immune genes, 'CXCR3' and 'CASP1' emerged as the highest contributors to PC1, with contributions of approximately 8.56 and 8.60 respectively. These genes might play crucial roles in discriminating between the different infection statuses. Their high contribution to the first principal component suggests that their expression is a significant source of variation in our dataset and potentially correlated with the infection types under study.

Other immune genes like 'IFNy', 'IL1RN', 'IRGM1', and 'MUC2' also showed significant contributions to PC1, with values ranging from approximately 6.12 to 7.89. The strong representation of these genes on PC1 implies that their expression varies significantly across the mice and infection types, and they might play a role in the different immune responses.

In contrast, some genes, including 'IL.13' and 'TICAM1', had relatively low contributions to PC1, approximately 0.04 and 0.92 respectively, suggesting that these genes may not vary as much across our samples, at least not in a way that corresponds to the main axis of variation represented by PC1.

It is important to note that the high or low contribution to PC1 doesn't necessarily imply biological importance or irrelevance of a given gene in response to infection. The PCA is a statistical tool that identifies patterns in the dataset, and the contributions are based on the variance of gene expression.

Further investigation is required to validate these findings and understand the biological implications, preferably through experimental validation and correlation with phenotypic outcomes."

The genes CASP1, CXCR3, MUC2, and IRGM1 are involved in various immune responses and have different roles:

CASP1 (Caspase 1): It plays a crucial role in the innate immune response by activating pro-inflammatory cytokines IL-1beta and IL-18. CASP1 is involved in initiating pyroptosis, a form of programmed cell death, and is associated with inflammatory conditions.

CXCR3 (C-X-C motif chemokine receptor 3): It is a chemokine receptor expressed on immune cells, including T cells and natural killer cells. CXCR3 is involved in the recruitment of immune cells to sites of inflammation and plays a role in immune responses against pathogens and tumors.

MUC2 (Mucin 2): It is a major component of the mucus layer that lines various epithelial surfaces, including the intestinal tract. MUC2 provides a physical barrier and helps protect against pathogens and other harmful substances.

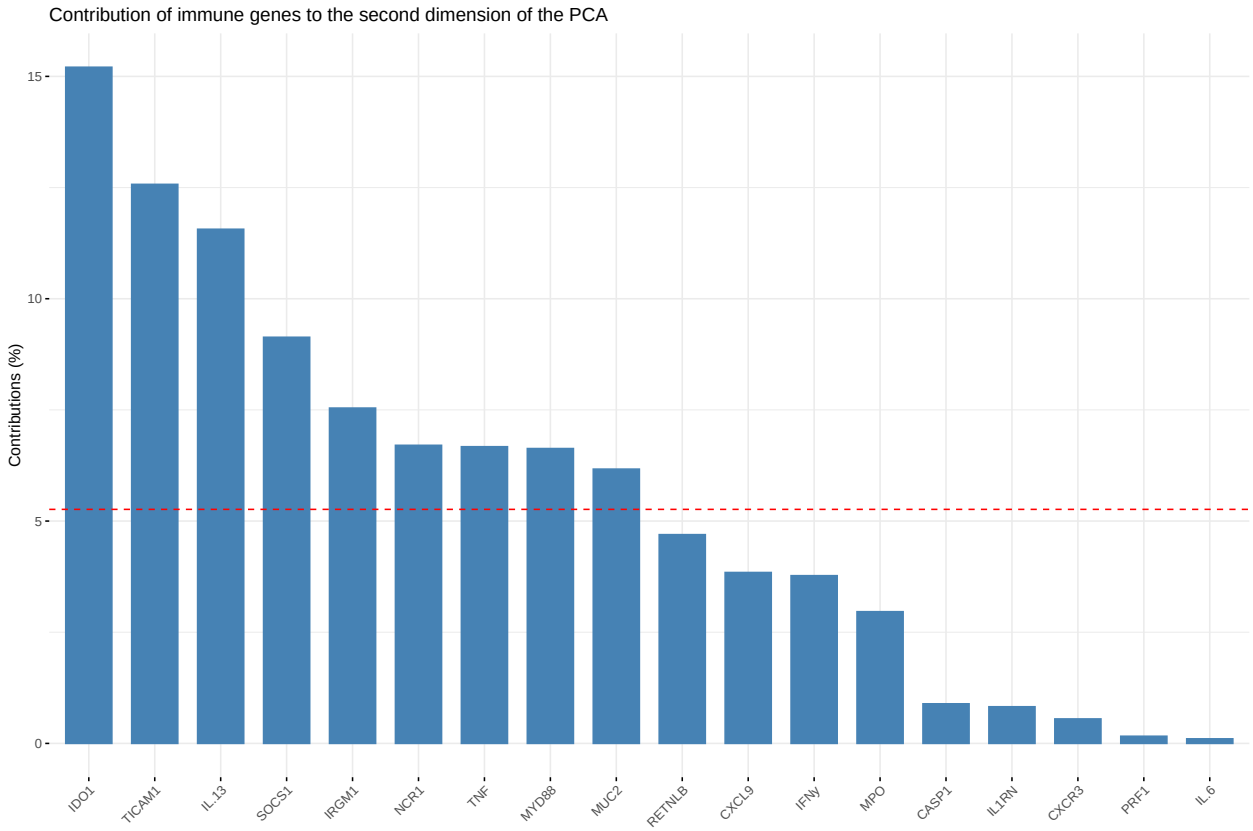
IRGM1 (Immunity-Related GTPase Family M Member 1): It is an immunity-related GTPase involved in host defense against intracellular pathogens. IRGM1 participates in autophagy, a cellular process that eliminates intracellular pathogens and helps regulate immune responses.

The common thread connecting these genes and immune responses is their involvement in the innate immune system. The innate immune system provides immediate defense against pathogens and triggers an inflammatory response. These genes and their respective proteins contribute to the recognition of pathogens, recruitment of immune cells, activation of cytokines, and protection of epithelial surfaces.

Regarding damaging processes to the host, excessive or dysregulated activation of immune responses can lead to tissue damage and inflammation. Inflammatory conditions associated with uncontrolled immune responses can have detrimental effects on the host. For example, chronic inflammation can contribute to tissue destruction, organ damage, and autoimmune diseases. Balancing the immune response is essential to prevent excessive damage while effectively combating pathogens or maintaining tissue homeostasis.

Contributions to the second dimension

```
# Contributions of variables to PC2
fviz_contrib(res.pca, choice = "var", axes = 2, top = 18,
             title = "Contribution of immune genes to the second dimension of the PCA")
```



Title: "Contribution of Immune Genes to the Second Principal Component of Gene Expression Analysis"

Description:

"The plot represents the contribution of each of the 19 immune genes to the second principal component (PC2) of a principal component analysis based on gene expression data derived from controlled infection experiments in laboratory mice.

The gene 'IL.13' shows the highest contribution to PC2, with a value of approximately 11.56. This significant contribution suggests that the expression level of this gene varies in a way that is significantly related to the second largest source of variation in our data set and could be strongly correlated with another aspect of the infection types that is different from what PC1 is capturing.

The gene 'IDO1' also exhibits a considerable contribution to PC2, with a value of around 15.21. The 'TICAM1' gene, on the other hand, shows a lower contribution to PC2, with a value of approximately 1.26, indicating that its variance might be less related to the second dimension of variation.

Notably, 'CXCR3' and 'CASP1', which were the highest contributors to PC1, show significantly lower contributions to PC2 (0.55 and 0.89 respectively), suggesting that the variations in their expression are less aligned with the axis represented by PC2.

It's essential to remember that these values represent statistical correlations and not necessarily biological significance. The actual biological interpretation of these findings will require further investigation, possibly involving a correlation with phenotypic outcomes or functional studies of these genes."

In the context of intestinal parasites, the involvement of IDO1, TICAM1, SOCS1, and Interleukin 13 (IL-13) can vary depending on the specific parasite and the host immune response. Here is some information regarding their roles in the context of intestinal parasites:

IDO1 (Indoleamine 2,3-dioxygenase 1): IDO1 is an enzyme involved in the metabolism of tryptophan, an essential amino acid. In the context of intestinal parasites, IDO1 expression can be induced as part of the

host immune response to control parasite infections. IDO1 can modulate immune responses by degrading tryptophan, leading to the inhibition of parasite growth and the generation of immunoregulatory metabolites.

TICAM1 (Toll-like receptor adaptor molecule 1): TICAM1 is an adaptor protein involved in Toll-like receptor (TLR) signaling pathways. TLRs play a crucial role in recognizing microbial components and initiating immune responses. In the context of intestinal parasites, TICAM1 may be involved in the activation of TLR signaling pathways in response to parasite-derived molecules, leading to the production of pro-inflammatory cytokines and the initiation of an immune response.

SOCS1 (Suppressor of Cytokine Signaling 1): SOCS1 is a negative regulator of cytokine signaling. It helps control the duration and intensity of cytokine signaling by inhibiting downstream signaling pathways. In the context of intestinal parasites, SOCS1 may be induced as part of a negative feedback mechanism to prevent excessive immune activation and inflammation caused by the host response to the parasite. It can inhibit cytokine signaling pathways, including those involving IL-13, to regulate the immune response.

Interleukin 13 (IL-13): IL-13 is an immunoregulatory cytokine that plays a role in mediating the host immune response to intestinal parasites. IL-13 can be produced by immune cells, such as T cells and innate lymphoid cells, in response to parasite infection. It can contribute to the activation of immune cells, promote the recruitment of inflammatory cells to the site of infection, and induce the production of mucus and other factors that can help expel parasites from the intestine.

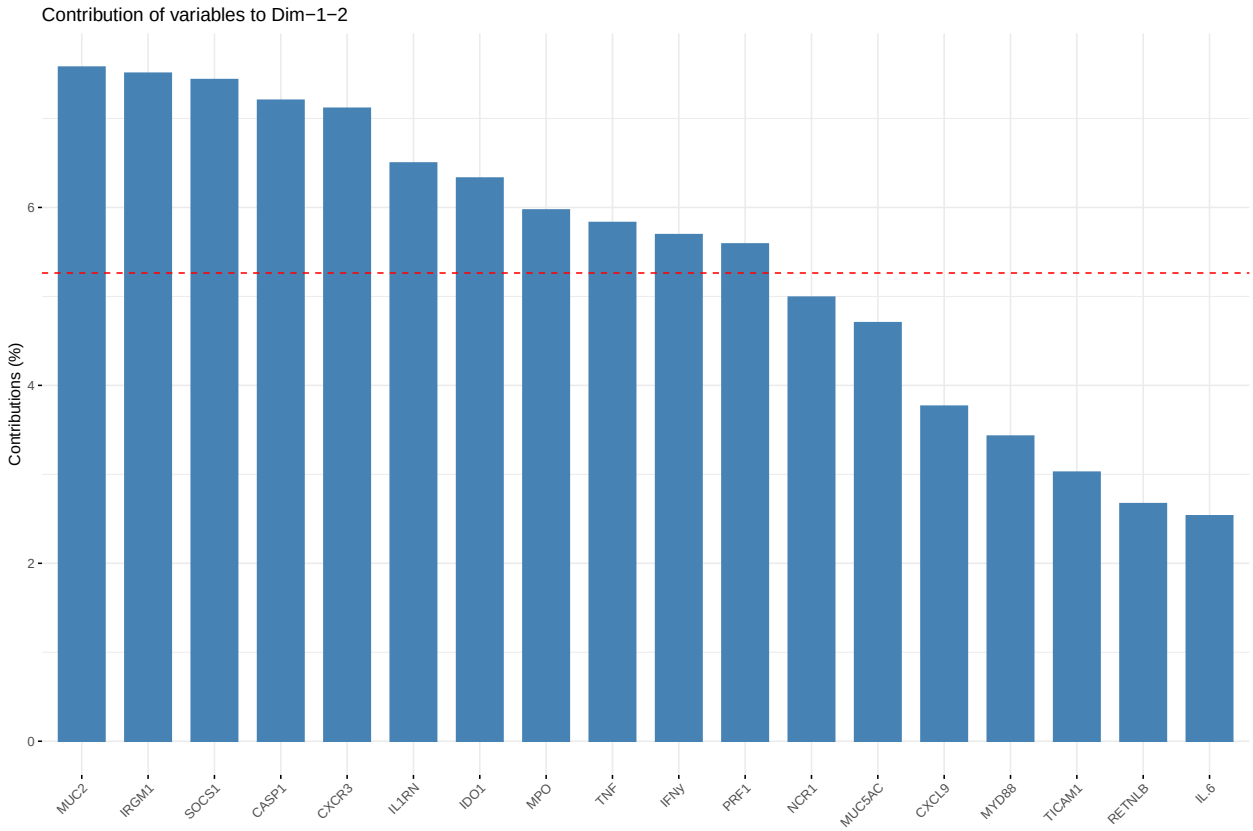
It's important to note that the specific roles and interactions of these factors can vary depending on the type of intestinal parasite, the host immune response, and the specific mechanisms of immune evasion employed by the parasite.

The reactions involving IDO1, TICAM1, SOCS1, and Interleukin 13 (IL-13) in response to intestinal parasites can have both beneficial and potentially damaging effects on the host. The degree of damage depends on various factors, including the type and virulence of the parasite, the intensity and duration of the immune response, and the susceptibility of the host.

Beneficial Effects: These reactions are part of the host's immune defense against parasites and aim to control and eliminate the infection. The immune response triggered by these factors can help limit parasite growth, prevent parasite dissemination, and promote the clearance of the parasites from the intestinal tract. Additionally, the production of cytokines, such as IL-13, can stimulate the production of mucus and enhance the barrier function of the intestinal epithelium, aiding in the expulsion of parasites.

Potential Damaging Effects: In some cases, the immune response can be excessive or dysregulated, leading to tissue damage and inflammation. Chronic or uncontrolled immune responses can cause collateral damage to the host tissues and disrupt normal physiological processes in the intestine. Excessive production of pro-inflammatory cytokines and chemokines can contribute to tissue damage, inflammation, and alteration of the gut microbiota. Additionally, prolonged activation of immune signaling pathways, such as those involving TICAM1 and SOCS1, can lead to immune system dysfunction and contribute to chronic inflammation.

It's important to note that the balance between the beneficial and damaging effects of the immune response is crucial. In an effective immune response, the host can effectively control and eliminate the parasites while minimizing damage to its own tissues. However, in some cases, the immune response can be insufficient, allowing the parasites to persist and cause chronic infections or tissue damage. The overall impact on the host depends on the interplay between the host immune response, the parasite's virulence factors, and the host's genetic and environmental factors.

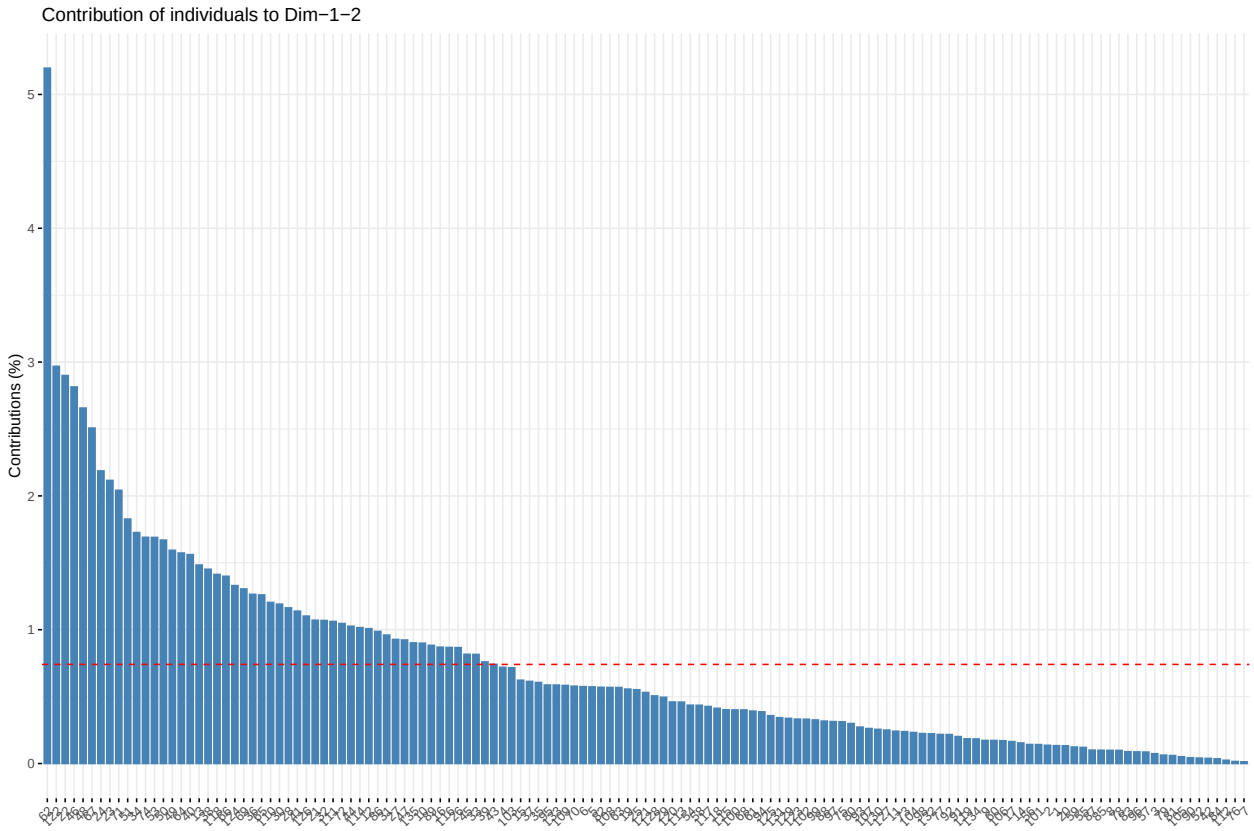


The red dashed line on the graph above indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would be 1/length(variables) = 1/10 = 10%. For a given component, a variable with a contribution larger than this cutoff could be considered as important in contributing to the component.

Note that, the total contribution of a given variable, on explaining the variations retained by two principal components, say PC1 and PC2, is calculated as contrib = [(C1 * Eig1) + (C2 * Eig2)]/(Eig1 + Eig2), where

C1 and C2 are the contributions of the variable on PC1 and PC2, respectively Eig1 and Eig2 are the eigenvalues of PC1 and PC2, respectively. Recall that eigenvalues measure the amount of variation retained by each PC. In this case, the expected average contribution (cutoff) is calculated as follow: As mentioned above, if the contributions of the 10 variables were uniform, the expected average contribution on a given PC would be 1/10 = 10%. The expected average contribution of a variable for PC1 and PC2 is : [(10* Eig1) + (10 * Eig2)]/(Eig1 + Eig2)

To visualize the contribution of individuals to the first two principal components:



PCA + Biplot combination

Figure description: "This PCA biplot presents a comprehensive view of our dataset, incorporating both the samples' positioning (mice) and variable contributions (immune gene expression). Each point represents an individual mouse categorized by infection status: Eimeria falciformis infection, Eimeria ferrisi infection, and uninfected. The two primary axes, PC1 and PC2, encapsulate the greatest variance in immune response as dictated by the measured immune gene expressions.

Vectors, or arrows, represent the 19 key immune genes analyzed. The direction and length of each arrow indicate how each gene contributes to the two principal components. The acute angle between the arrows corresponds to the correlation between the genes: a small angle indicates a strong positive correlation, a large angle suggests a weak correlation, and an orthogonal angle represents no correlation.

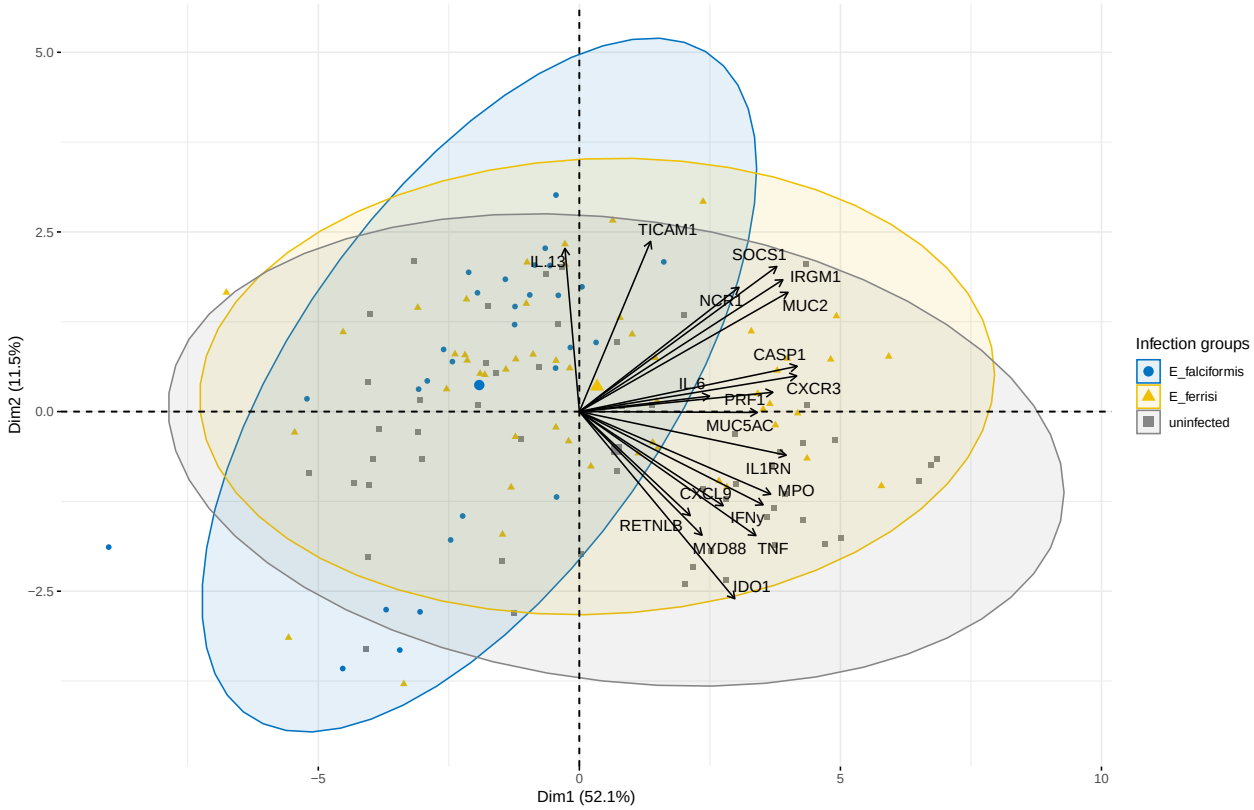
Observing the cluster patterns of the different infection statuses, we notice mice infected with Eimeria falciformis aggregate towards one section of the biplot, and similarly for the mice infected with Eimeria ferrisi and the uninfected mice. This clustering illustrates the distinct expression patterns of immune genes under the three conditions.

Importantly, the proximity of each group to the various gene vectors provides an initial indication of which genes may be driving these distinct patterns. For instance, if mice infected with Eimeria falciformis are near vectors corresponding to certain immune genes, this suggests these genes may be particularly upregulated in response to Eimeria falciformis infection, and so on for each group.

Overall, the biplot provides an integrative view of the immune responses in the European house mouse under different infection statuses, offering key insights into the genes potentially pivotal in mediating these responses."

```
#select same rows in the first table
lab <- lab[row.names(genes), ]

fviz_pca_biplot(res.pca,
```

```
              col.ind = lab$Parasite_challenge, palette = "jco",
              addEllipses = TRUE, label = "var",
              col.var = "black", repel = TRUE,
              legend.title = "Infection groups",
              title = "")
```



In the following example, we want to color both individuals and variables by groups. The trick is to use pointshape = 21 for individual points. This particular point shape can be filled by a color using the argument fill.ind. The border line color of individual points is set to "black" using col.ind. To color variable by groups, the argument col.var will be used.

Linear models: Predicting weight loss with the PCA eigenvectors

Are the pc1 and pc2 components relevant in predicting the maximum weight loss per mouse?

```
##
## Call:
## lm(formula = WL_max ~ pc1 + pc2, data = lab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7713  -4.2625   0.6397   4.8968  16.8997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.6962     0.5844 -16.591  < 2e-16 ***
## pc1           0.7126     0.1857   3.838 0.000191 ***
## pc2          -2.3593     0.3955  -5.965 2.11e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6.79 on 132 degrees of freedom
## Multiple R-squared:  0.276,  Adjusted R-squared:  0.265
## F-statistic: 25.16 on 2 and 132 DF,  p-value: 5.538e-10
```

```
## [1] 905.2669
```

Our linear regression model aimed to predict maximum weight loss (WL_max) in laboratory mice based on the first two principal components of immune gene expression (pc1 and pc2). The model accounted for approximately 27.6% of the variability in weight loss (Multiple R-squared = 0.276), with an adjusted R-squared of 0.265. The adjusted R-squared considers the number of predictors in the model, offering a more conservative estimate of the model's explanatory power.

The F-statistic is 25.16 with a very low p-value (5.538e-10), implying that at least one of the predictors is significantly related to the outcome variable.

Specifically, both pc1 and pc2 are significantly related to the outcome variable (WL_max). For pc1, an increase of 1 unit is associated with an average increase of 0.7126 units in WL_max, with a standard error of 0.1857 (t = 3.838, p = 0.000191). For pc2, an increase of 1 unit is associated with an average decrease of 2.3593 units in WL_max, indicating a negative relationship between pc2 and weight loss (standard error = 0.3955, t = -5.965, p = 2.11e-08).

The residual standard error of 6.79 suggests there's some variation in weight loss that our model doesn't explain, indicating potential room for model improvement. The intercept term (-9.6962) represents the predicted weight loss when pc1 and pc2 are both zero, though this may not have a meaningful interpretation in the context of this study.

```
##
## Call:
## lm(formula = WL_max ~ pc1 + pc2 + Parasite_challenge, data = lab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1978  -3.6326   0.1512   4.6046  19.0753
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -15.4210     1.2141 -12.702  < 2e-16 ***
## pc1                            0.4024     0.1797   2.239   0.0269 *
## pc2                           -2.0155     0.3793  -5.314 4.53e-07 ***
## Parasite_challengeE_ferrisi    6.7836     1.4954   4.536 1.29e-05 ***
## Parasite_challengeuninfected   7.7796     1.5501   5.019 1.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.21 on 130 degrees of freedom
## Multiple R-squared:  0.4036, Adjusted R-squared:  0.3853
## F-statistic:    22 on 4 and 130 DF,  p-value: 6.967e-14
```

```
## [1] 883.0803
```

In the second linear regression model, we expanded our predictor set to include the first two principal components of immune gene expression (pc1 and pc2) and the experimental infection group (Parasite_challenge). This model had a better fit to our data, explaining approximately 40.36% of the variability in maximum weight loss (WL_max) in laboratory mice (Multiple R-squared = 0.4036), with an adjusted R-squared of 0.3853.

The F-statistic of this model is 22, and the corresponding p-value (6.967e-14) is very small, providing strong

evidence that at least one of the predictors is significantly related to the outcome variable.

Looking at individual predictors:

For pc1, an increase of 1 unit is associated with an average increase of 0.4024 units in WL_max, with a standard error of 0.1797 (t = 2.239, p = 0.0269).

For pc2, an increase of 1 unit is associated with an average decrease of 2.0155 units in WL_max, indicating a negative relationship between pc2 and weight loss (standard error = 0.3793, t = -5.314, p = 4.53e-07).

Compared to the baseline category of Parasite_challenge (E_falciformis), mice in the E_ferrisi group had an average increase of 6.7836 units in WL_max (standard error = 1.4954, t = 4.536, p = 1.29e-05), and those in the uninfected group had an average increase of 7.7796 units (standard error = 1.5501, t = 5.019, p = 1.67e-06).

The residual standard error of this model is 6.21, showing an improvement over the first model. This suggests the inclusion of Parasite_challenge as a predictor helps account for more variation in the outcome. The intercept term (-15.4210) represents the predicted weight loss when all predictors are zero (which may not be meaningful in this context).

Lastly, the Akaike Information Criterion (AIC) of this model is 883.0803. The AIC is a measure of the relative quality of statistical models for a given dataset; lower AIC values indicate a better-fitting model. Thus, this model would be preferred over any other models with a higher AIC such as the model before.

```
##
## Call:
## lm(formula = WL_max ~ pc1 + pc2 + hybrid_status, data = lab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7296  -4.5289   0.4619   5.0095  17.8237
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -8.5522     1.0294  -8.308 1.26e-13 ***
## pc1                               0.5505     0.2407   2.287   0.0238 *
## pc2                              -2.0370     0.4910  -4.149 6.07e-05 ***
## hybrid_statusF0 M. m. musculus   -3.2572     1.5805  -2.061   0.0414 *
## hybrid_statusF1 hybrid            1.6874     2.1959   0.768   0.4437
## hybrid_statusF1 M. m. domesticus -2.6981     2.8962  -0.932   0.3533
## hybrid_statusF1 M. m. musculus    1.8058     3.3404   0.541   0.5897
## hybrid_statusother               -1.7817     1.9248  -0.926   0.3564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.713 on 127 degrees of freedom
## Multiple R-squared:  0.3192, Adjusted R-squared:  0.2817
## F-statistic: 8.507 on 7 and 127 DF,  p-value: 1.611e-08
```

```
## [1] 906.9537
```

In this third model, the predictors include the first two principal components of immune gene expression (pc1 and pc2) and the hybrid_status, which indicates the type of hybrid mouse. The model explains approximately 31.92% of the variability in maximum weight loss (WL_max) in laboratory mice (Multiple R-squared = 0.3192), with an adjusted R-squared of 0.2817.

The F-statistic of this model is 8.507, and the corresponding p-value (1.611e-08) is very small, providing strong evidence that at least one of the predictors is significantly related to the outcome variable.

Examining individual predictors:

For pc1, an increase of 1 unit is associated with an average increase of 0.5505 units in WL_max, with a standard error of 0.2407 (t = 2.287, p = 0.0238).

For pc2, an increase of 1 unit is associated with an average decrease of 2.0370 units in WL_max, indicating a negative relationship between pc2 and weight loss (standard error = 0.4910, t = -4.149, p = 6.07e-05).

The hybrid_status categories are compared against the baseline (presumably "F0 M. m. domesticus"). The "F0 M. m. musculus" group had a lower average weight loss by 3.2572 units (standard error = 1.5805, t = -2.061, p = 0.0414). The remaining hybrid groups did not exhibit a statistically significant difference from the baseline group.

The residual standard error of this model is 6.713, and the intercept term (-8.5522) represents the predicted weight loss when all predictors are zero (which may not be meaningful in this context).

Finally, the Akaike Information Criterion (AIC) of this model is 906.9537. The AIC is a measure of the relative quality of statistical models for a given dataset. Lower AIC values indicate a better-fitting model. However, in this case, the AIC is higher than the previous model, suggesting that the second model (including pc1, pc2, and Parasite_challenge) might provide a better fit to the data.

Try instead: LLR test (likelihood ration) (LM4 package)?

https://www.rdocumentation.org/packages/lmtest/versions/0.9-38/topics/lrtest

In this way you compare each model, with the different variables usesd to predict.

Another way is to compare the AIC. (function : step)

```
# Compare
llr_test <- anova(model_1_pc1_pc2, model_2_pc1_pc2_challenge)
print(llr_test)
```

```
## Analysis of Variance Table
##
## Model 1: WL_max ~ pc1 + pc2
## Model 2: WL_max ~ pc1 + pc2 + Parasite_challenge
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    132 6086.4
## 2    130 5013.3  2    1073.2 13.914 3.344e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

"In order to assess the predictive capacity of principal component scores (pc1 and pc2) and parasite challenge group on maximum weight loss (WL_max), two linear regression models were compared. The first model only included the principal component scores (pc1 and pc2), while the second model additionally accounted for the effect of the parasite challenge group.

A significant improvement in the prediction of maximum weight loss was observed when the parasite challenge group was included in the model ($F(2, 130) = 13.914$, $p < 0.00001$). This suggests that the parasite challenge group has a significant effect on the maximum weight loss and improves the predictive power of the model beyond that provided by the principal component scores alone.

Consequently, the second model (WL_max ~ pc1 + pc2 + Parasite_challenge) demonstrated superior explanatory power for maximum weight loss during infection experiments compared to the first model (WL_max ~ pc1 + pc2)."

```
# model_2_pc1_pc2_challenge <- lm(WL_max ~ pc1 + pc2 + Parasite_challenge, data = lab)

weight_no_pc1 <- lm(WL_max ~ pc2 + Parasite_challenge, data = lab)
weight_no_pc2 <- lm(WL_max ~ pc1  + Parasite_challenge, data = lab)
```

```
weight_no_Parasite_challenge <- lm(WL_max ~ pc1 + pc2, data = lab)
lrtest(model_2_pc1_pc2_challenge, weight_no_pc1)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc1 + pc2 + Parasite_challenge
## Model 2: WL_max ~ pc2 + Parasite_challenge
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   6 -435.54
## 2   5 -438.09 -1 5.108    0.02382 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(model_2_pc1_pc2_challenge, weight_no_pc2)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc1 + pc2 + Parasite_challenge
## Model 2: WL_max ~ pc1 + Parasite_challenge
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -435.54
## 2   5 -448.81 -1 26.535  2.588e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(model_2_pc1_pc2_challenge, weight_no_Parasite_challenge)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc1 + pc2 + Parasite_challenge
## Model 2: WL_max ~ pc1 + pc2
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -435.54
## 2   4 -448.63 -2 26.187  2.059e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(weight_no_pc1, weight_no_pc2)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc2 + Parasite_challenge
## Model 2: WL_max ~ pc1 + Parasite_challenge
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -438.09
## 2   5 -448.81  0 21.427  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

"To investigate the importance of principal component scores (pc1 and pc2) and parasite challenge group in predicting maximum weight loss (WL_max) in mice during infection experiments, we performed likelihood ratio tests comparing different linear regression models.

The first comparison was between a model with both principal component scores and parasite challenge group (WL_max ~ pc1 + pc2 + Parasite_challenge) and a model excluding pc1 (WL_max ~ pc2 + Parasite_challenge). The results indicated that the inclusion of pc1 significantly improved model fit (deltax^2(1) = 5.108, p = 0.0238).

The second comparison was between the model with both principal component scores and parasite challenge group (WL_max ~ pc1 + pc2 + Parasite_challenge) and a model excluding pc2 (WL_max ~ pc1 + Parasite_challenge). The model with pc2 was significantly better at explaining the variation in maximum weight loss ($deltax^2(1) = 26.535$, $p < 0.0001$).

The third comparison was between the model with both principal component scores and parasite challenge group (WL_max ~ pc1 + pc2 + Parasite_challenge) and a model excluding the parasite challenge group (WL_max ~ pc1 + pc2). The inclusion of the parasite challenge group significantly improved the model fit ($deltax^2(2) = 26.187$, $p < 0.0001$).

The final comparison was between a model with pc2 and parasite challenge group (WL_max ~ pc2 + Parasite_challenge) and a model with pc1 and parasite challenge group (WL_max ~ pc1 + Parasite_challenge). The model with pc2 was significantly better than the model with pc1 in explaining the variation in maximum weight loss ($deltax^2(0) = 21.427$, $p < 0.0001$).

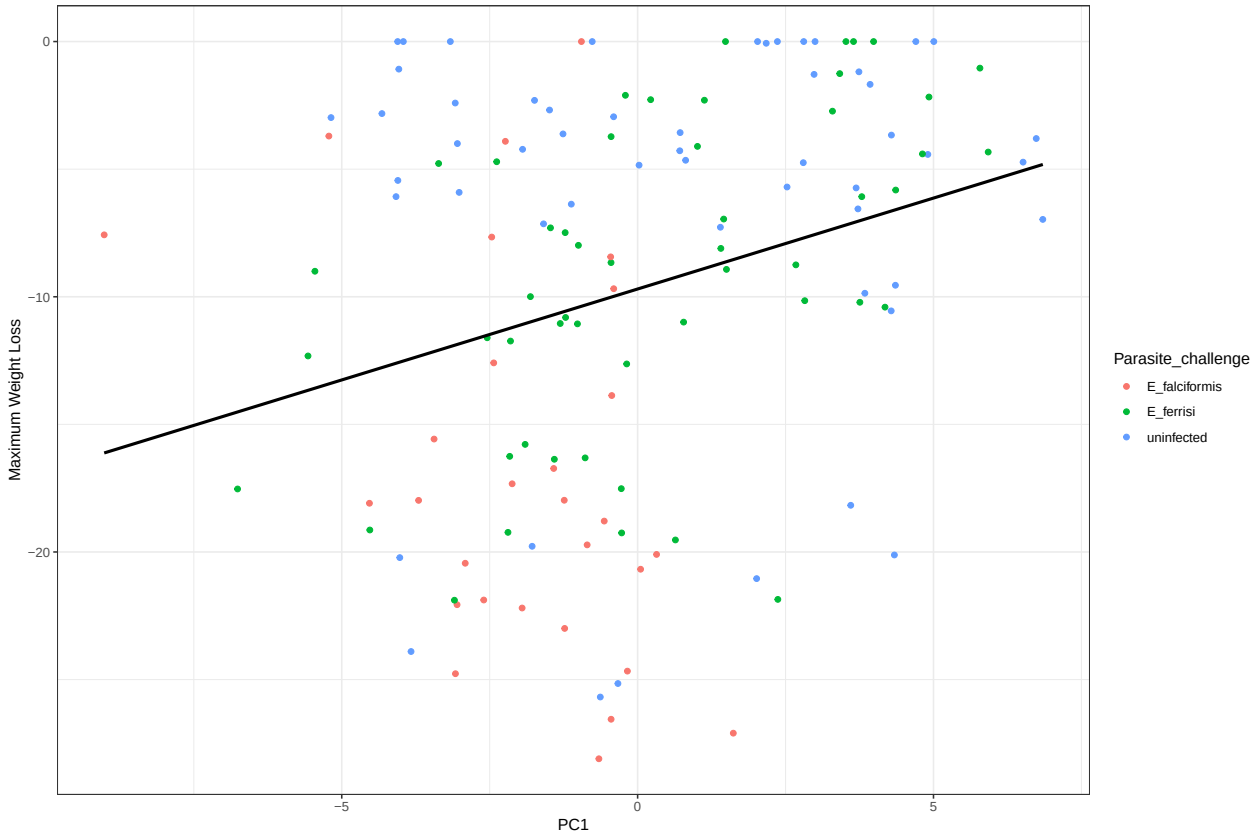
These results suggest that both principal component scores and the parasite challenge group contribute significantly to the predictive capacity of the model for maximum weight loss in mice during infection experiments. The full model including both principal component scores and the parasite challenge group (WL_max ~ pc1 + pc2 + Parasite_challenge) provides the best fit to the data."

Visualizing the regression models

scatter plot with regression lines

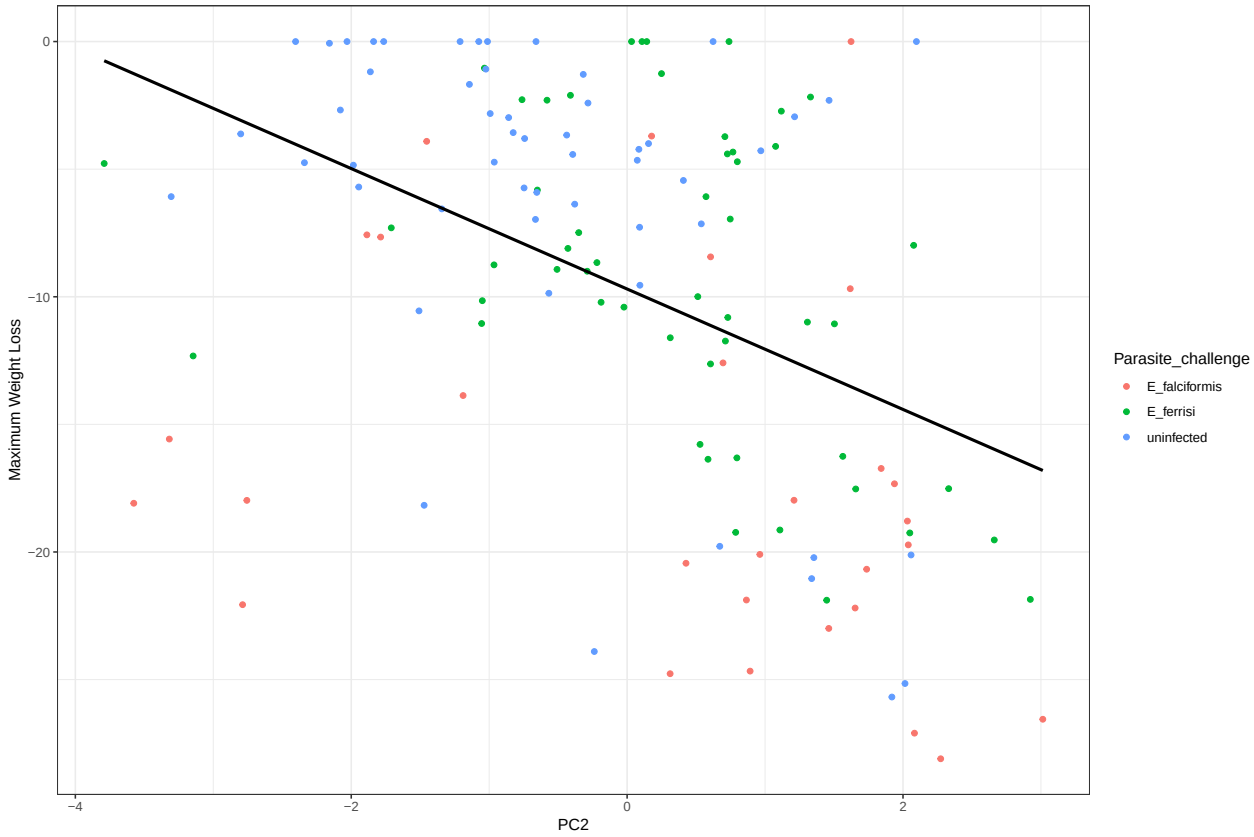
```
# for the model with pc1 and pc2
ggplot(lab, aes(x = pc1, y = WL_max)) +
  geom_point(aes(color = Parasite_challenge)) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(x = "PC1", y = "Maximum Weight Loss") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# for the model with pc2 and Parasite_challenge
ggplot(lab, aes(x = pc2, y = WL_max)) +
  geom_point(aes(color = Parasite_challenge)) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(x = "PC2", y = "Maximum Weight Loss") +
  theme_bw()
```

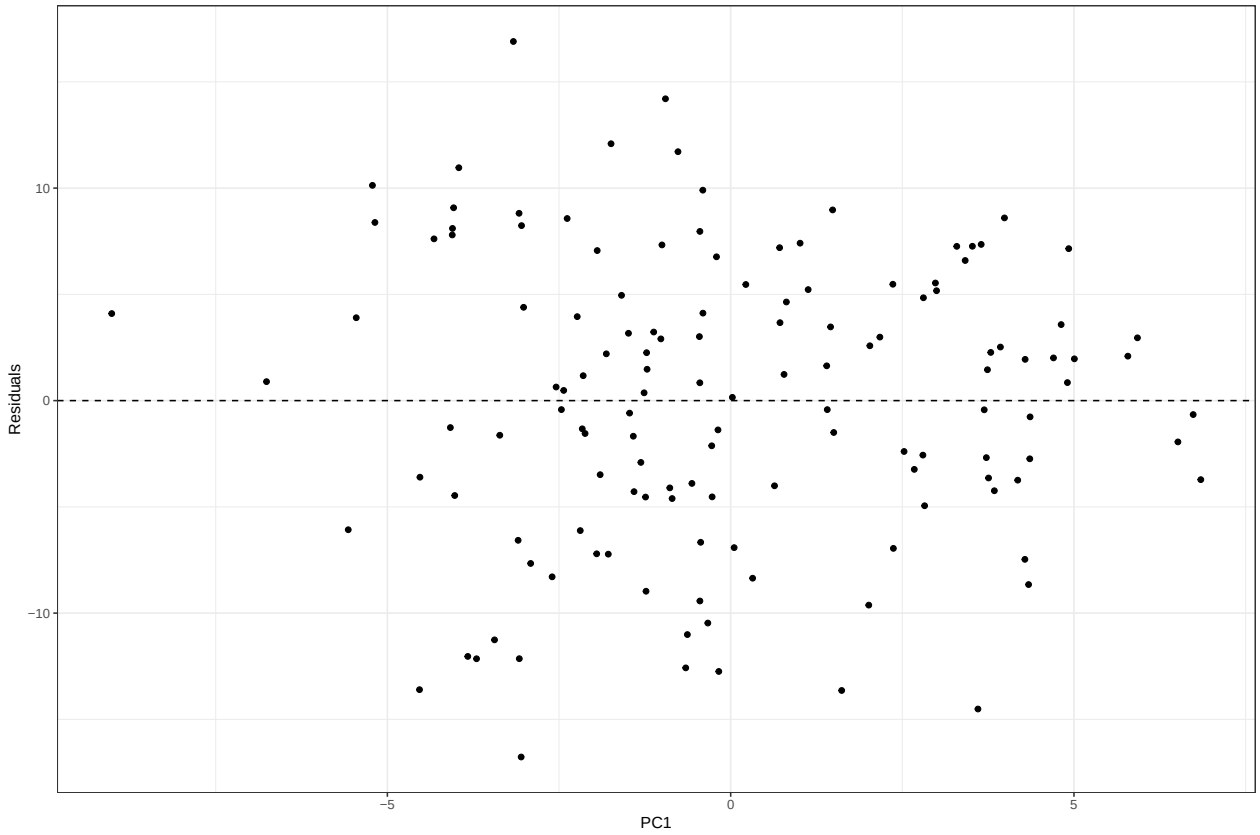
```
## `geom_smooth()` using formula = 'y ~ x'
```



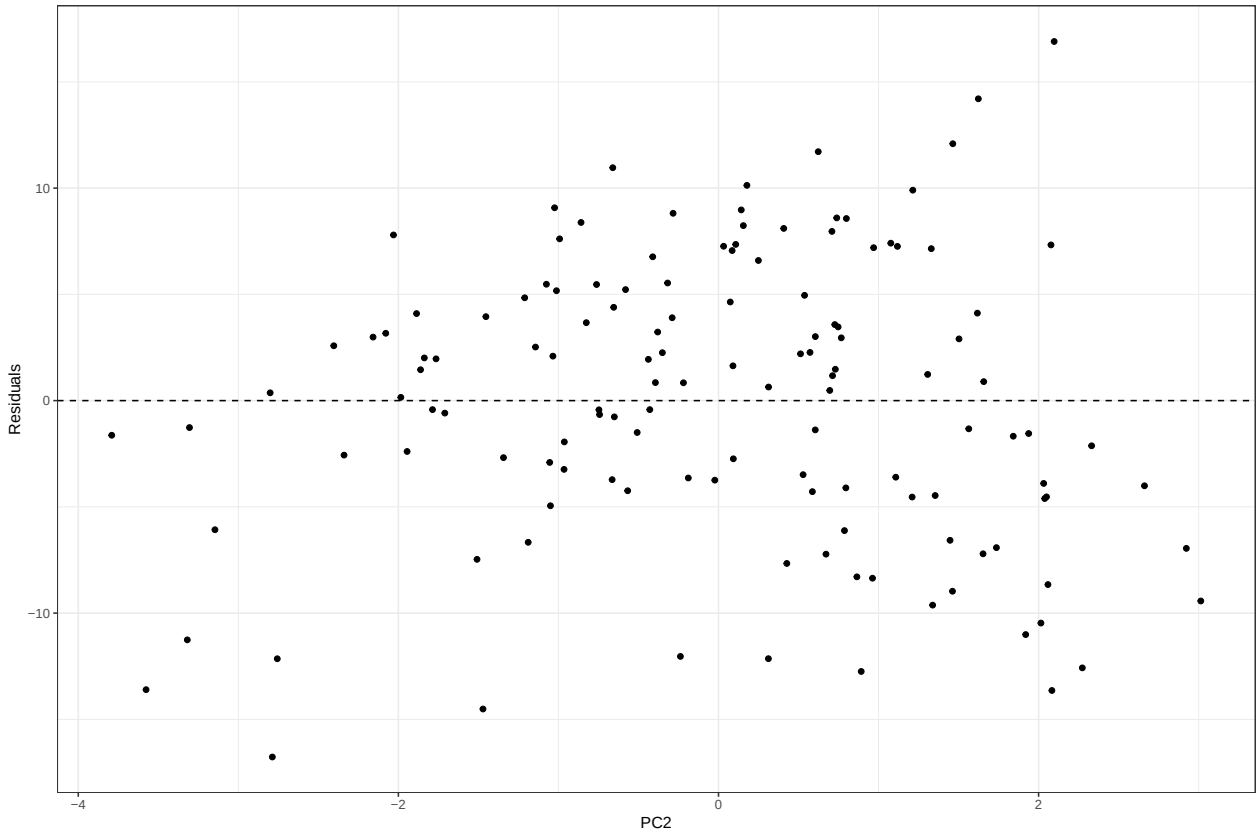
Residual Plots

Residual plots help to assess if the residuals have constant variance (homoscedasticity), which is an important assumption in linear regression. You can plot the residuals against the fitted values, or against each predictor variable.

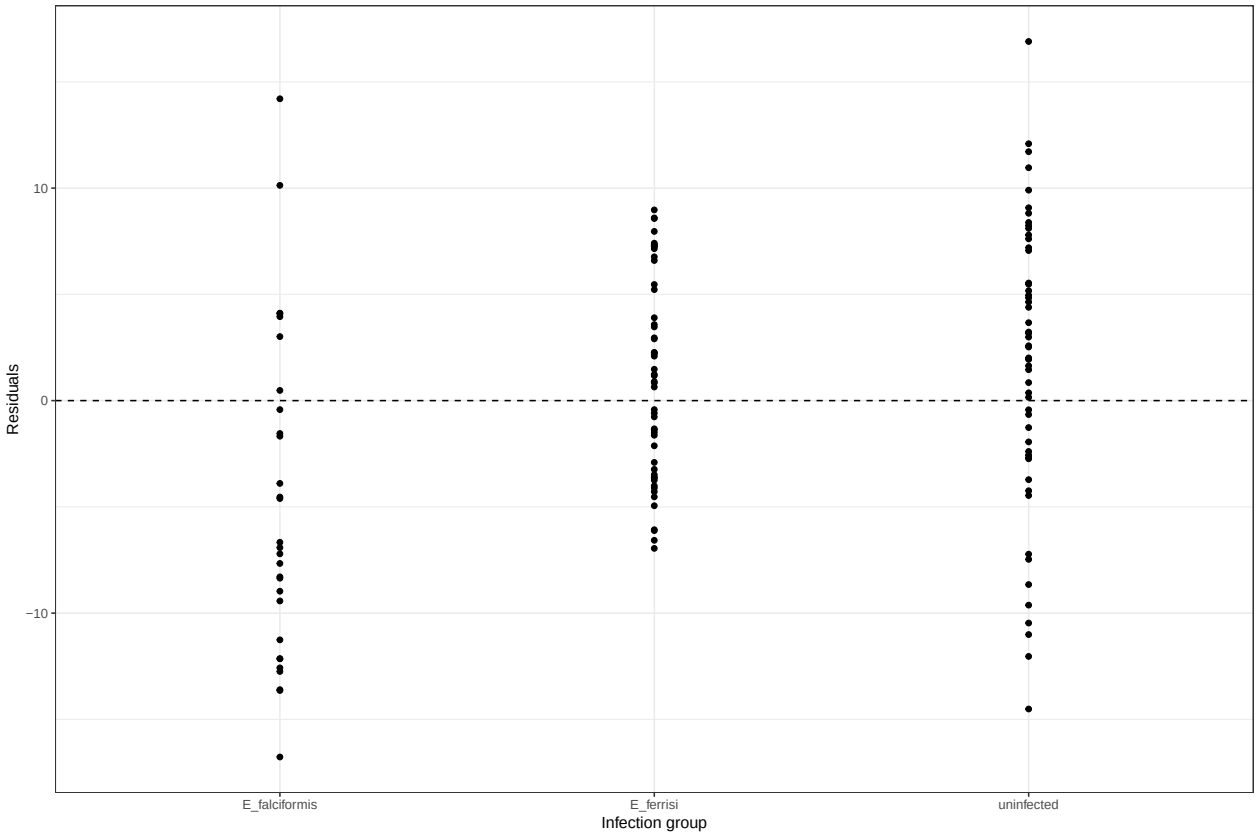
```
# calculate residuals for the model with pc1 and pc2
lab$residuals_pc1_pc2 <- resid(model_1_pc1_pc2)

ggplot(lab, aes(x = pc1, y = residuals_pc1_pc2)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "PC1", y = "Residuals") +
  theme_bw()
```




```
ggplot(lab, aes(x = pc2, y = residuals_pc1_pc2)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "PC2", y = "Residuals") +
  theme_bw()
```



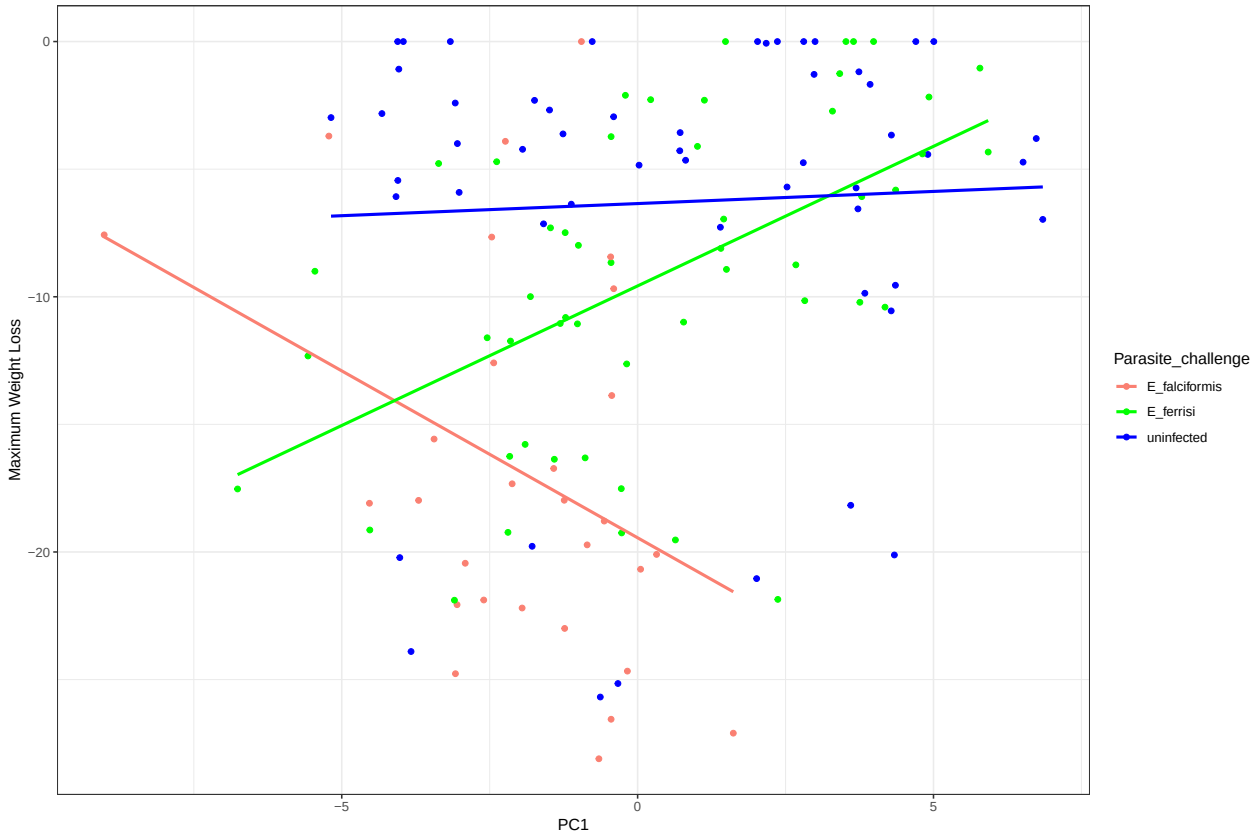

```
ggplot(lab, aes(x = Parasite_challenge, y = residuals_pc1_pc2)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Infection group", y = "Residuals") +
  theme_bw()
```



3D plots

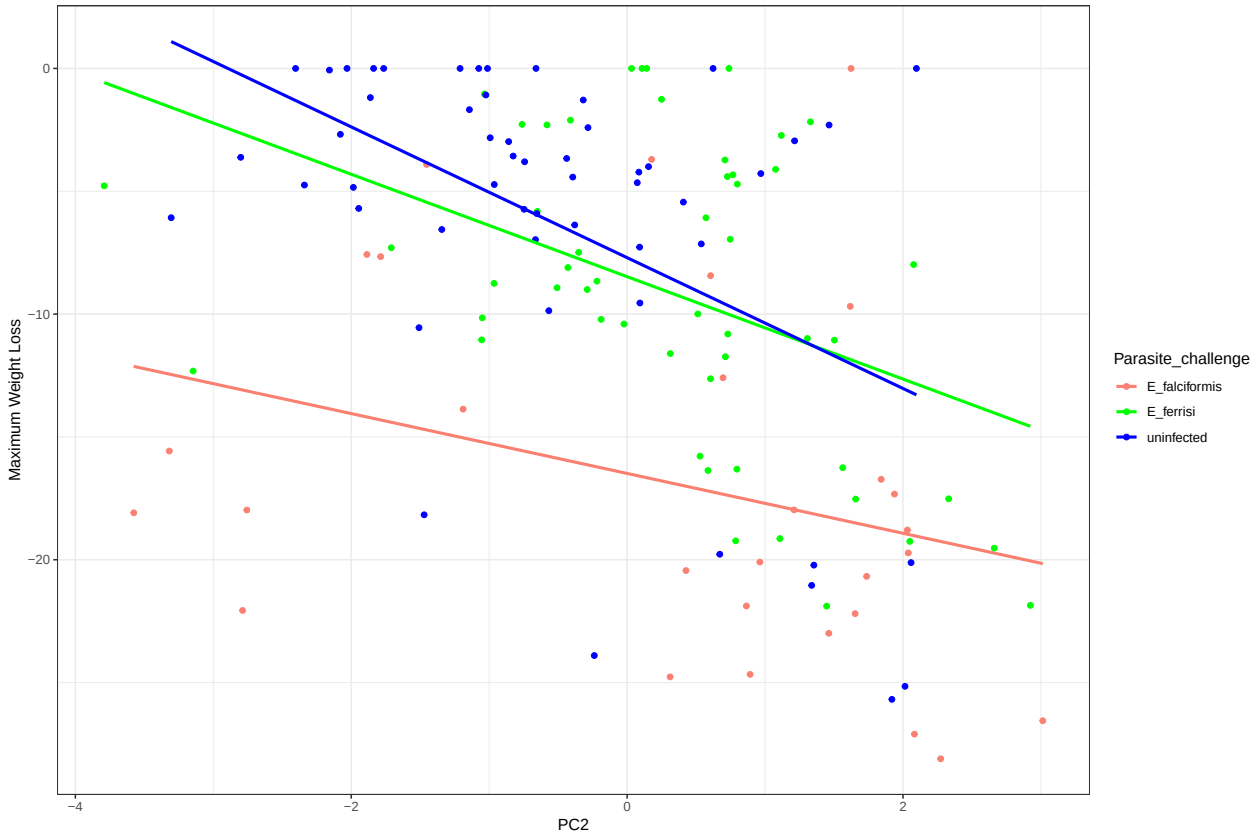
```
# First, make sure Parasite_challenge is a factor
lab$Parasite_challenge <- as.factor(lab$Parasite_challenge)

# Then, define the color for each level of Parasite_challenge
color_mapping <- c("E_falciformis" = "salmon",
                   "E_ferrisi" = "green",
                   "uninfected" = "blue")

# Now create the scatter plot using this color mapping
ggplot(lab, aes(x = pc1, y = WL_max, color = Parasite_challenge)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, aes(color = Parasite_challenge)) +
  scale_color_manual(values = color_mapping) +
  labs(x = "PC1", y = "Maximum Weight Loss") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

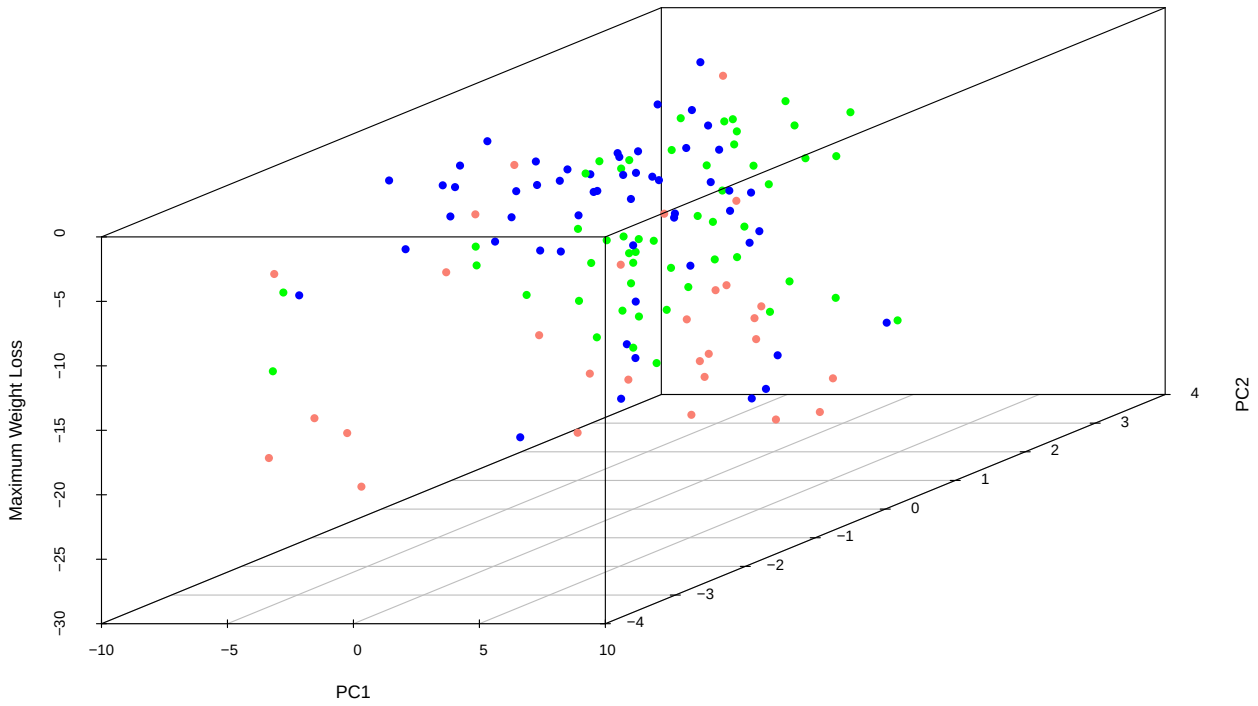



```
# Now create the scatter plot using this color mapping
ggplot(lab, aes(x = pc2, y = WL_max, color = Parasite_challenge)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, aes(color = Parasite_challenge)) +
  scale_color_manual(values = color_mapping) +
  labs(x = "PC2", y = "Maximum Weight Loss") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
library(scatterplot3d)


# Create a new color column in your dataframe by mapping 'Parasite_challenge' to your colors
lab$color <- color_mapping[lab$Parasite_challenge]

# 3D scatter plot
scatterplot3d(lab$pc1, lab$pc2, lab$WL_max, pch = 16, color = lab$color,
              xlab = "PC1", ylab = "PC2", zlab = "Maximum Weight Loss")
```



Heatmap

repeating the heatmap on the now imputed data

```
# turn the data frame into a matrix and transpose it. We want to have each cell
# type as a row name
gene <- t(as.matrix(gene))

# turn the first row into column names
gene %>%
    row_to_names(row_number = 1) -> heatmap_data

heatmap_data <- as.data.frame(heatmap_data)

table(rowSums(is.na(heatmap_data)) == nrow(heatmap_data))
```

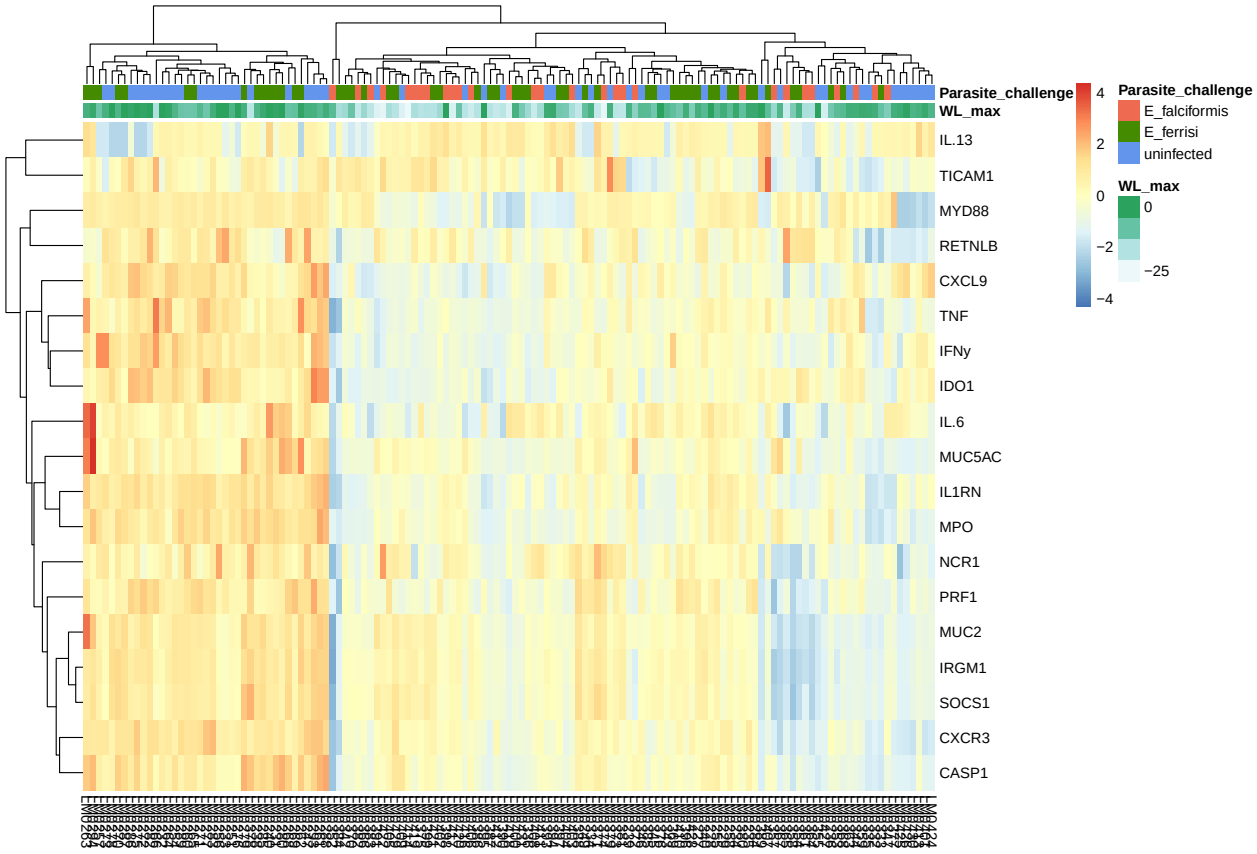
```
##
## FALSE
##    19
```

```
# turn the columns to numeric other wise the heatmap function will not work
heatmap_data[] <- lapply(heatmap_data, function(x) as.numeric(as.character(x)))

# remove columns with only NAs
heatmap_data <- Filter(function(x)!all(is.na(x)), heatmap_data)

#remove rows with only Nas
heatmap_data <-  heatmap_data[, colSums(is.na(heatmap_data)) !=
                                nrow(heatmap_data)]
```

```r
#Prepare the annotation data frame
annotation_df <- as_tibble(lab) %>%
    dplyr::select(c("Mouse_ID",  "WL_max", "Parasite_challenge"))

annotation_df <- unique(annotation_df)

annotation_df <- as.data.frame(annotation_df)




### Prepare the annotation columns for the heatmap
rownames(annotation_df) <- annotation_df$Mouse_ID


# Match the row names to the heatmap data frame
rownames(annotation_df) <- colnames(heatmap_data)

#remove the unecessary column
annotation_df <- annotation_df %>% dplyr::select(-Mouse_ID, )
```

Heatmap on gene expression data:

```r
# Define colors for each parasite
parasite_colors <- c("E_falciformis" = "coral2",
                     "E_ferrisi" = "chartreuse4",
                     "uninfected" = "cornflowerblue")

# Generate the heat map
pheatmap(heatmap_data, annotation_col = annotation_df,  scale = "row",
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         annotation_colors = list(Parasite_challenge = parasite_colors)) # use annotation_colors
```


Parasite_challenge
WL_max
IL.13
TICAM1
MYD88
RETNLB
CXCL9
TNF
IFNy
IDO1
IL.6
MUC5AC
IL1RN
MPO
NCR1
PRF1
MUC2
IRGM1
SOCS1
CXCR3
CASP1
4
2
0
−2
−4
Parasite_challenge
E_falciformis
E_ferrisi
uninfected
WL_max
0
−25