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

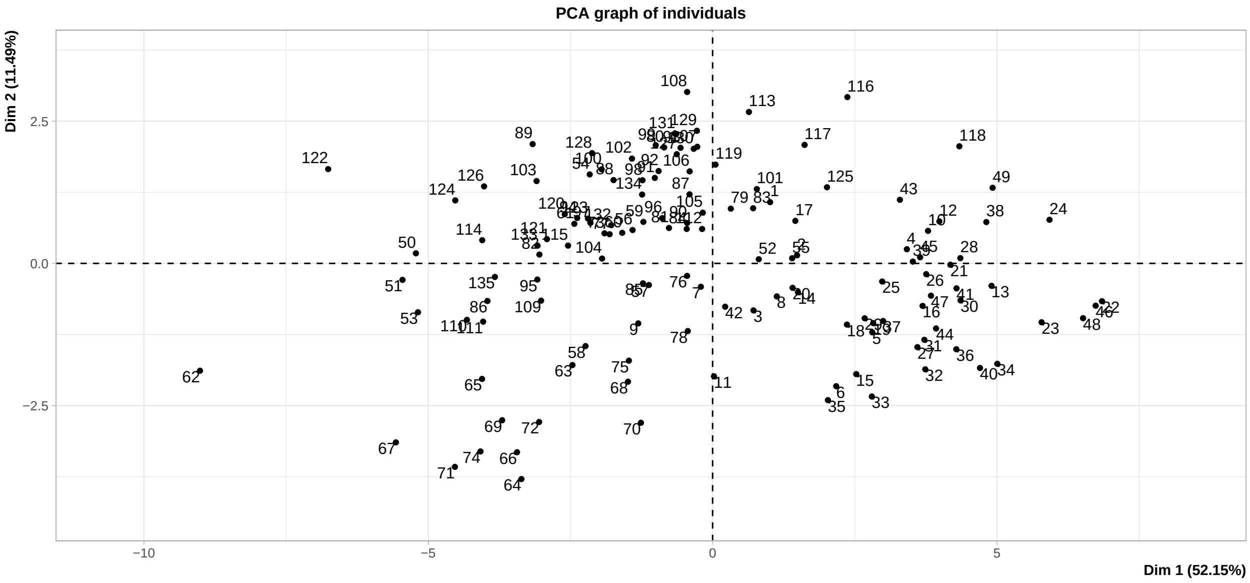
```
Gene_lab   <- c("IFNy", "CXCR3", "IL.6", "IL.13",
                "IL1RN","CASP1", "CXCL9", "IDO1", "IRGM1", "MPO",
                "MUC2", "MUC5AC", "MYD88", "NCR1", "PRF1", "RETNLB", "SOCS1",
                "TICAM1", "TNF") # "IL.12", "IRG6")
```

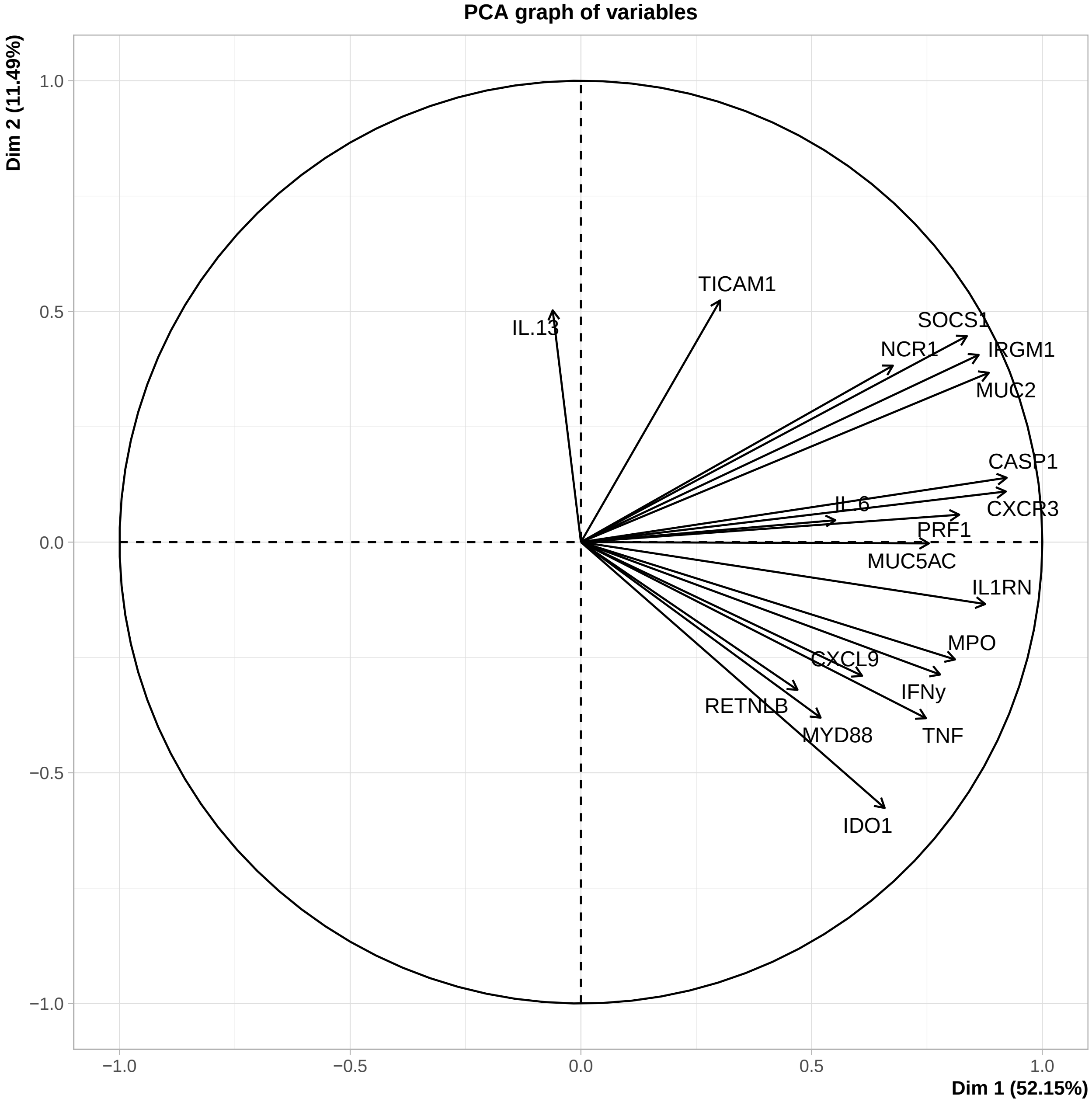
PCA on the lab genes

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)))
```

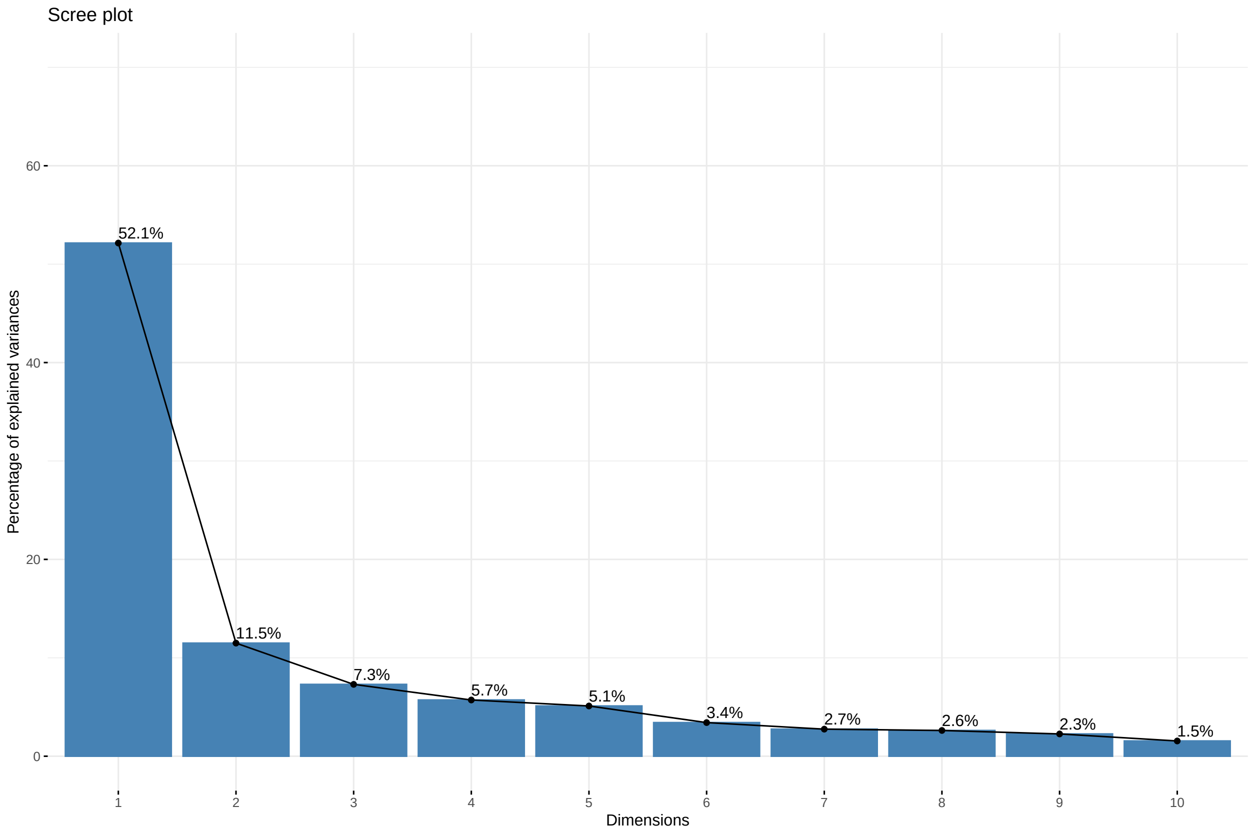
```
genes <- unique(gene)

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?

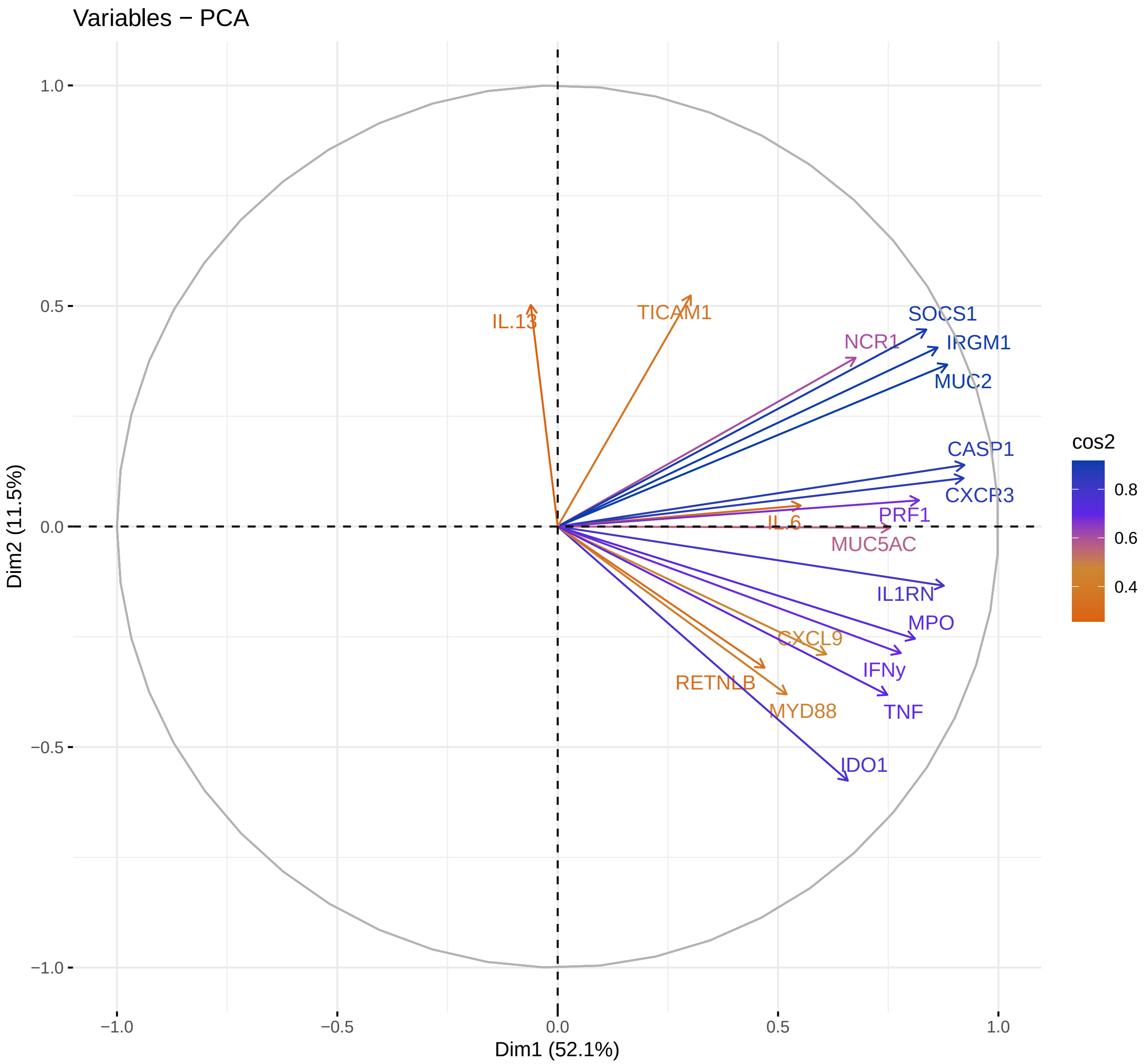
```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 70))
```



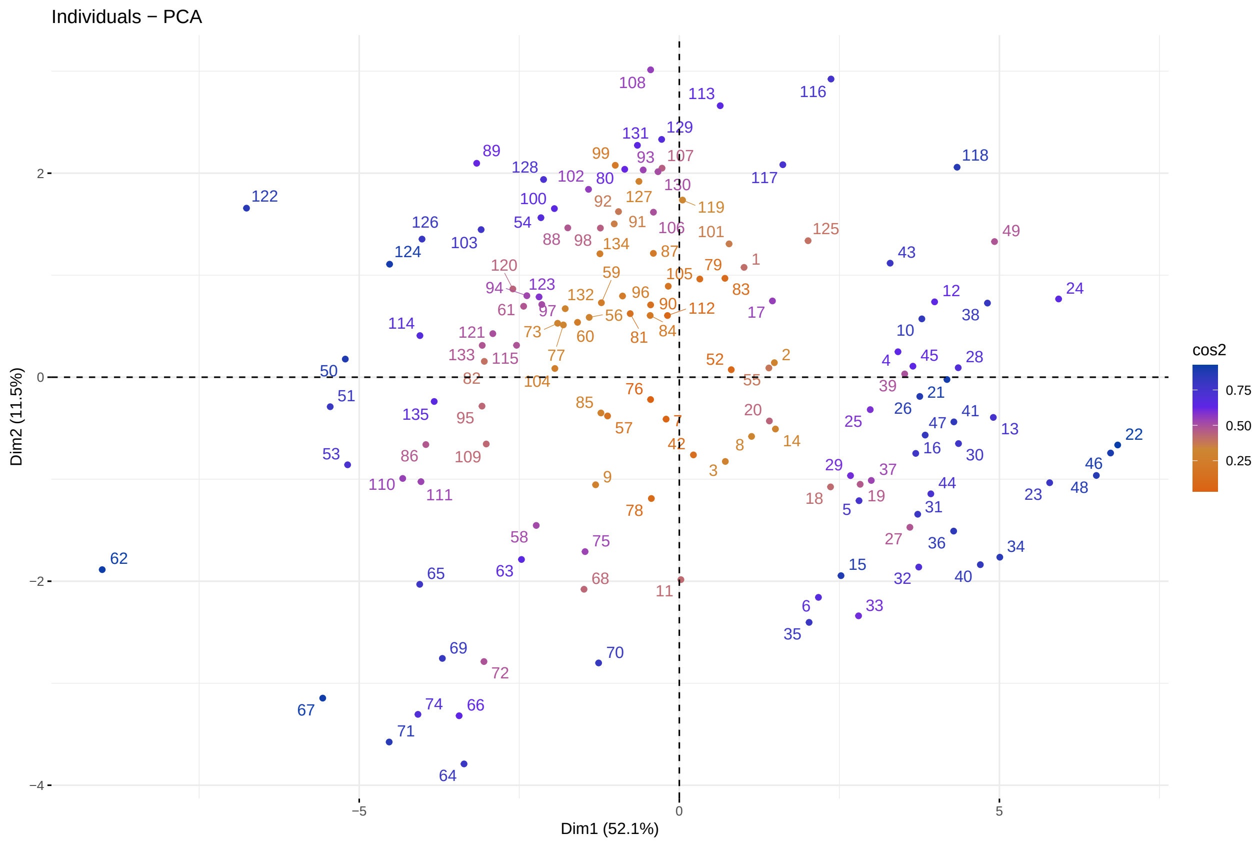
“Principal Component Analysis of Immune Gene Expression in Laboratory-Controlled Eimeria Infections”

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

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



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



Caution: When imputing data, the percentages of inertia associated with the first dimensions will be overestimated.

Another problem: the imputed data are, when the pca is performed considered like real observations. But they are estimations!!

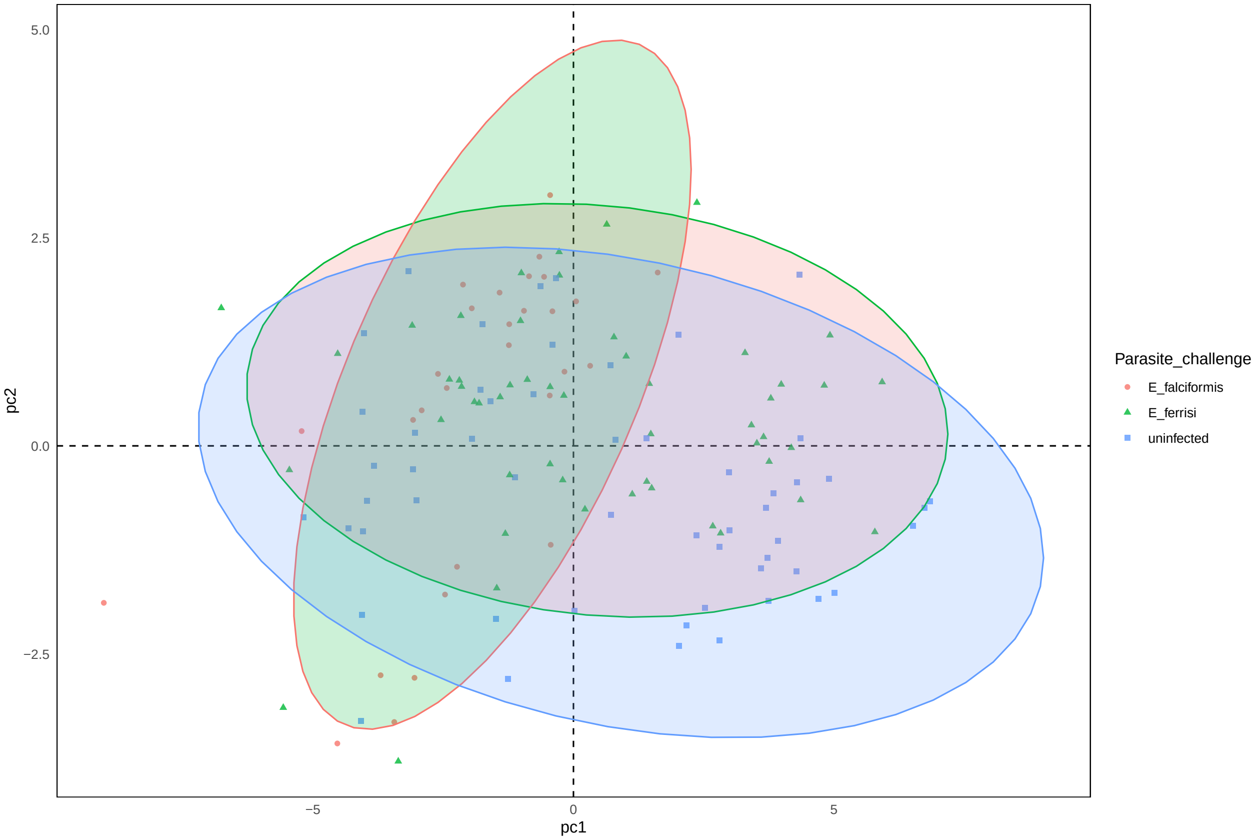
Visualizing uncertainty due to issing data:

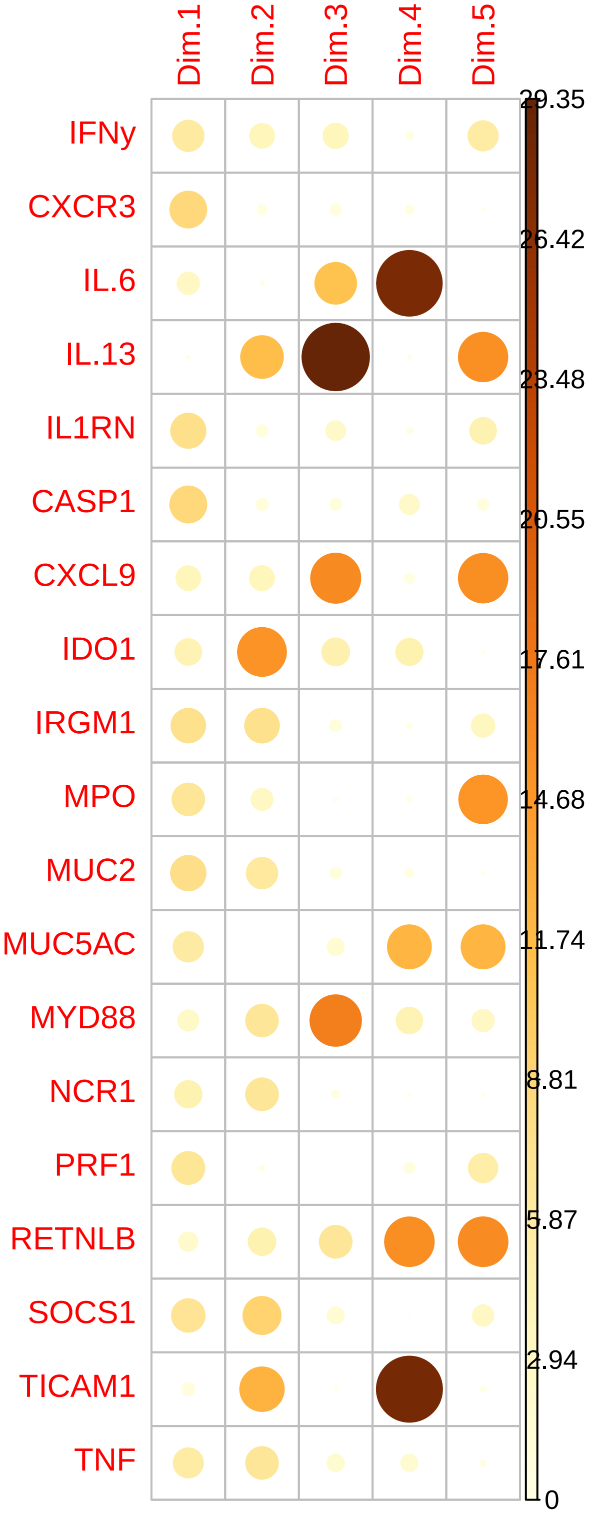
–> mulrimple imputation: generate several plausible values for each missing data point

We here visualize the variability, that is uncertainty on the plane defined by two pca axes.

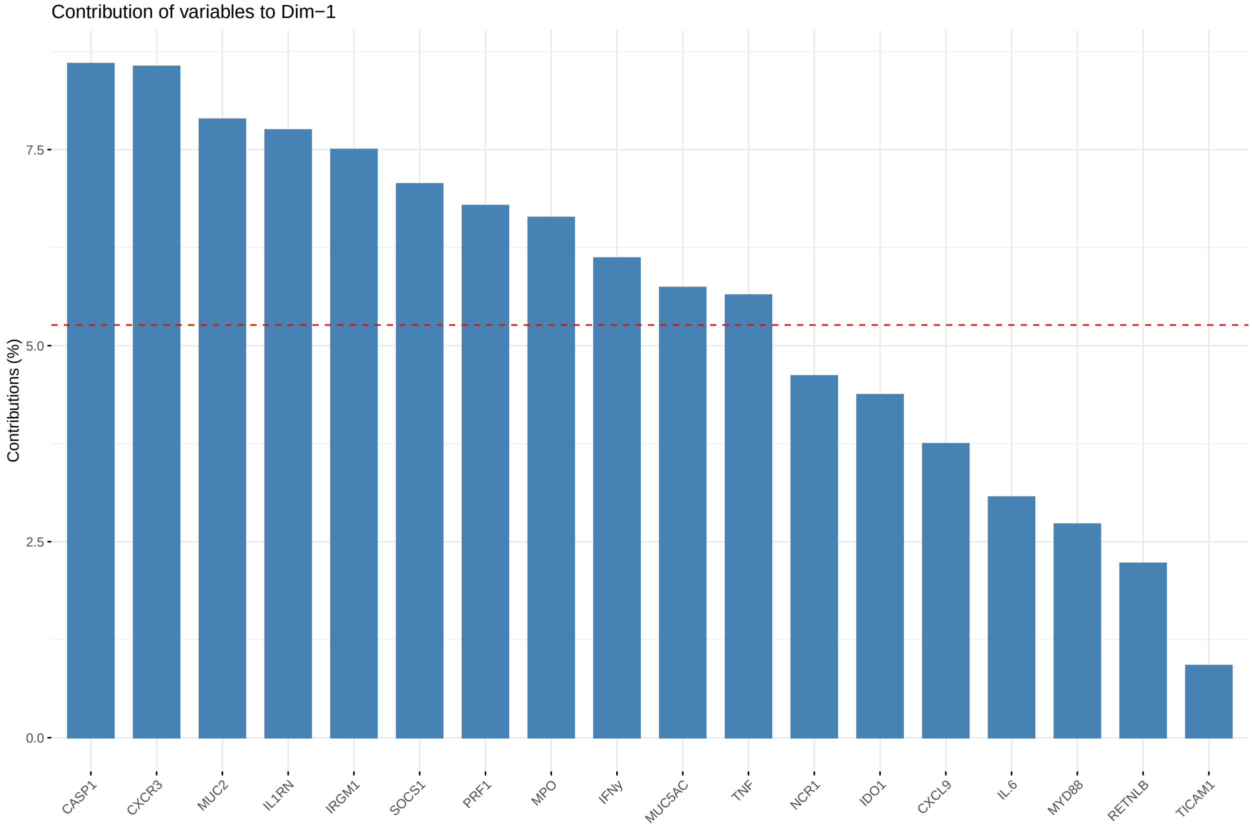
Biplot of the imputed gene pca

```
#Now we can make our initial plot of the PCA.
lab %>%
  ggplot(aes(x = pc1, y = pc2,
             color = Parasite_challenge,
             shape = Parasite_challenge)) +
  geom_hline(yintercept = 0, lty = 2) +
  geom_vline(xintercept = 0, lty = 2) +
  geom_point(alpha = 0.8) +
  stat_ellipse(geom="polygon",
               aes(fill = challenge_infection),
               alpha = 0.2, show.legend = FALSE,
               level = 0.95) +
  theme_minimal() +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"))
```


5.0
2.5
0.0
−2.5
pc2
−5
0
5
pc1
Parasite_challenge
● E_falciformis
▲ E_ferrisi
■ uninfected



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:



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.

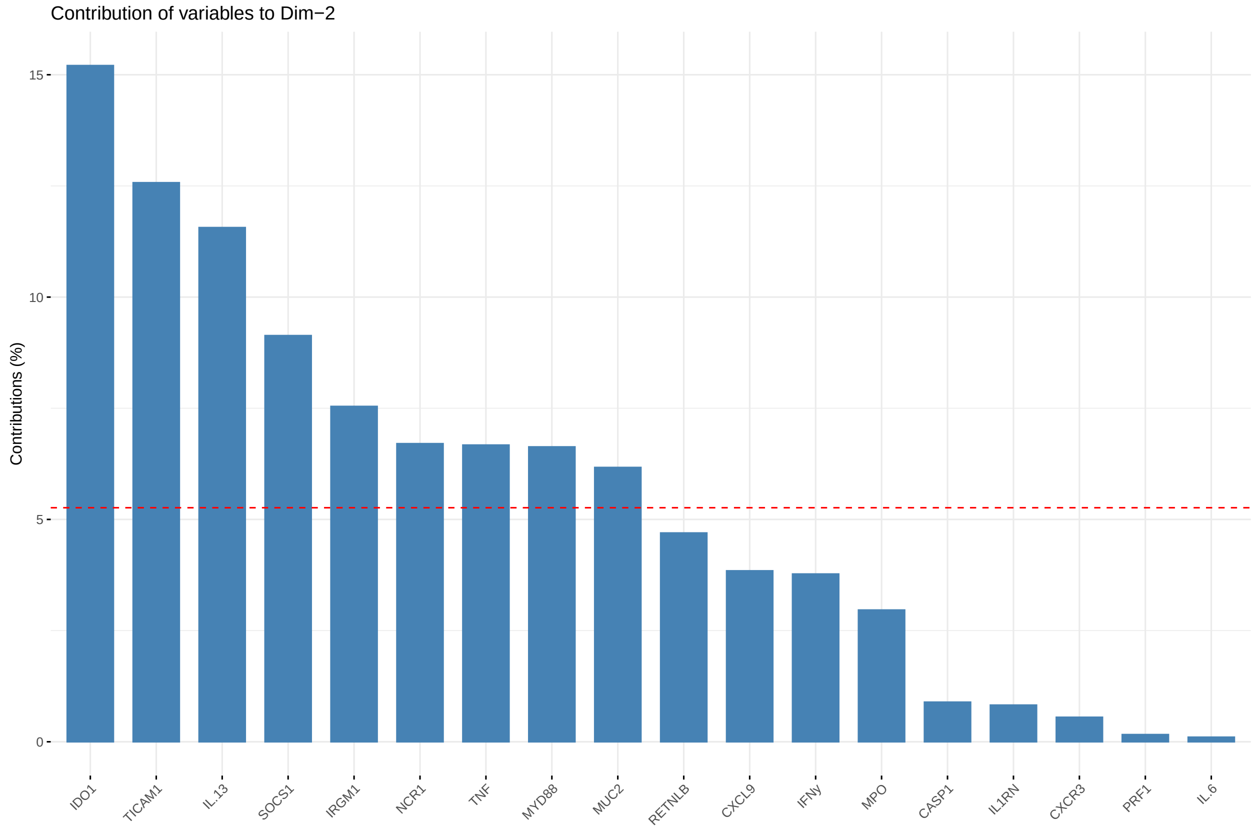
MUC2 (Mucin 2): It is a major component of the mucus layer that lines various epithelial surfaces, including the intestinal tract. MUC2 provides a physical barrier and helps protect against pathogens and other harmful substances.

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.

```r
# Contributions of variables to PC2
fviz_contrib(res.pca, choice = "var", axes = 2, top = 18)
```



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.

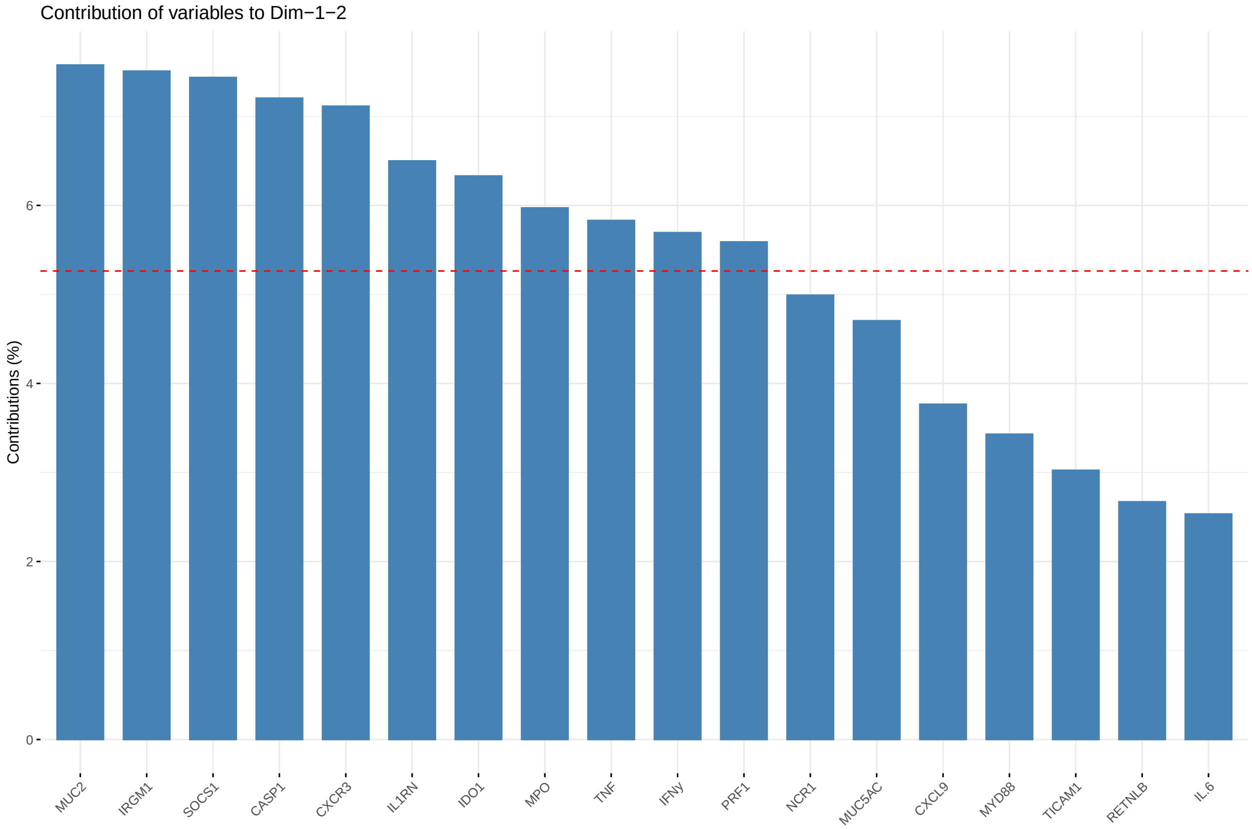
It's important to note that the specific roles and interactions of these factors can vary depending on the type of intestinal parasite, the host immune response, and the specific mechanisms of immune evasion employed by the parasite.

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.

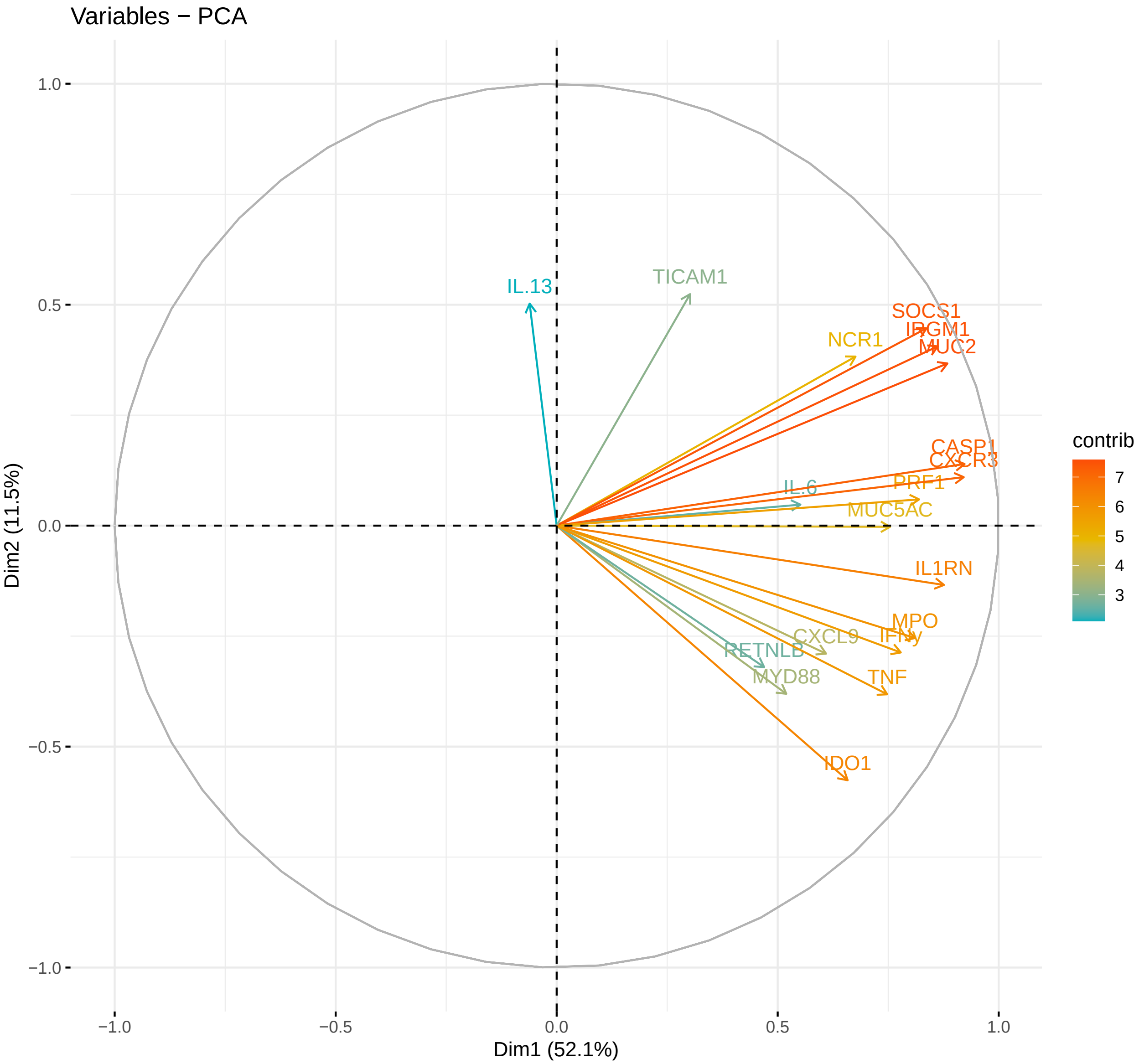
It's important to note that the balance between the beneficial and damaging effects of the immune response is crucial. In an effective immune response, the host can effectively control and eliminate the parasites while minimizing damage to its own tissues. However, in some cases, the immune response 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.



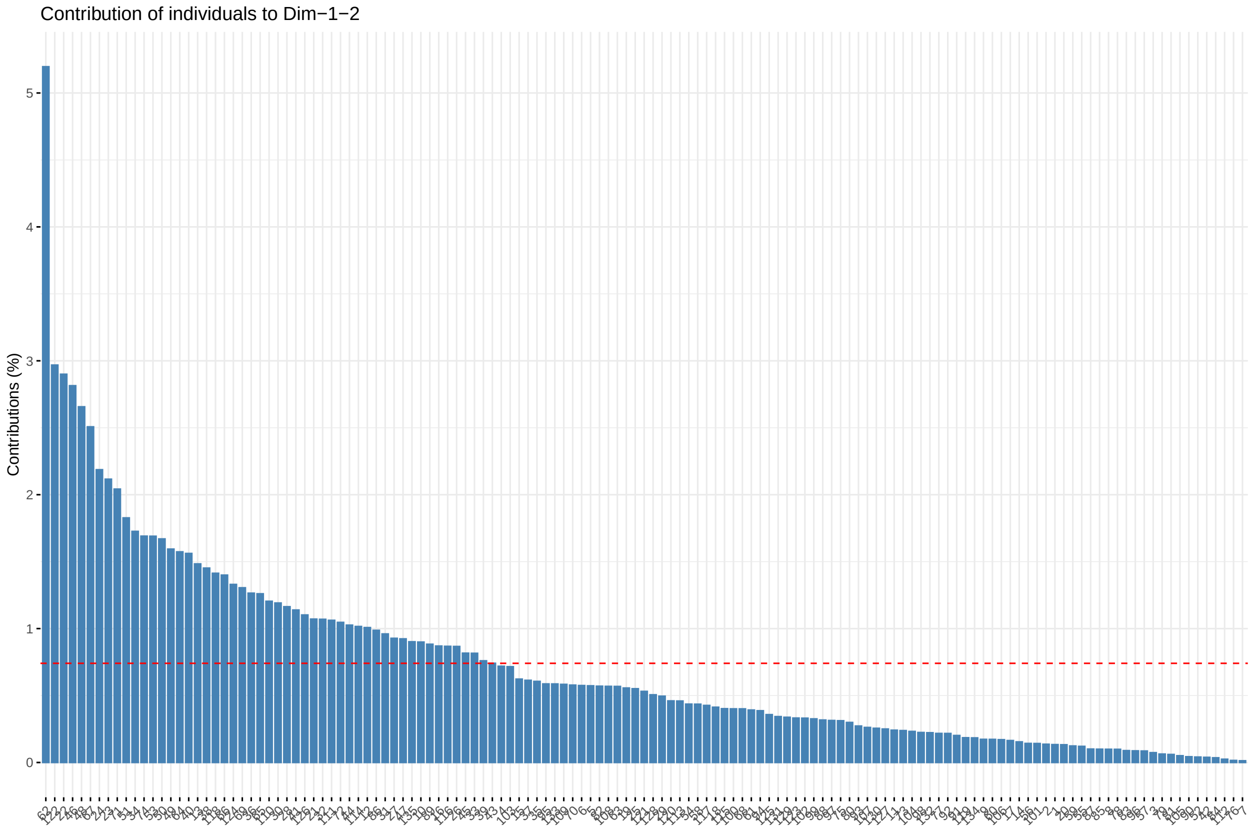
The red dashed line on the graph above indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would be 1/length(variables) = 1/10 = 10%. For a given component, a variable with a contribution larger than this cutoff could be considered as important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

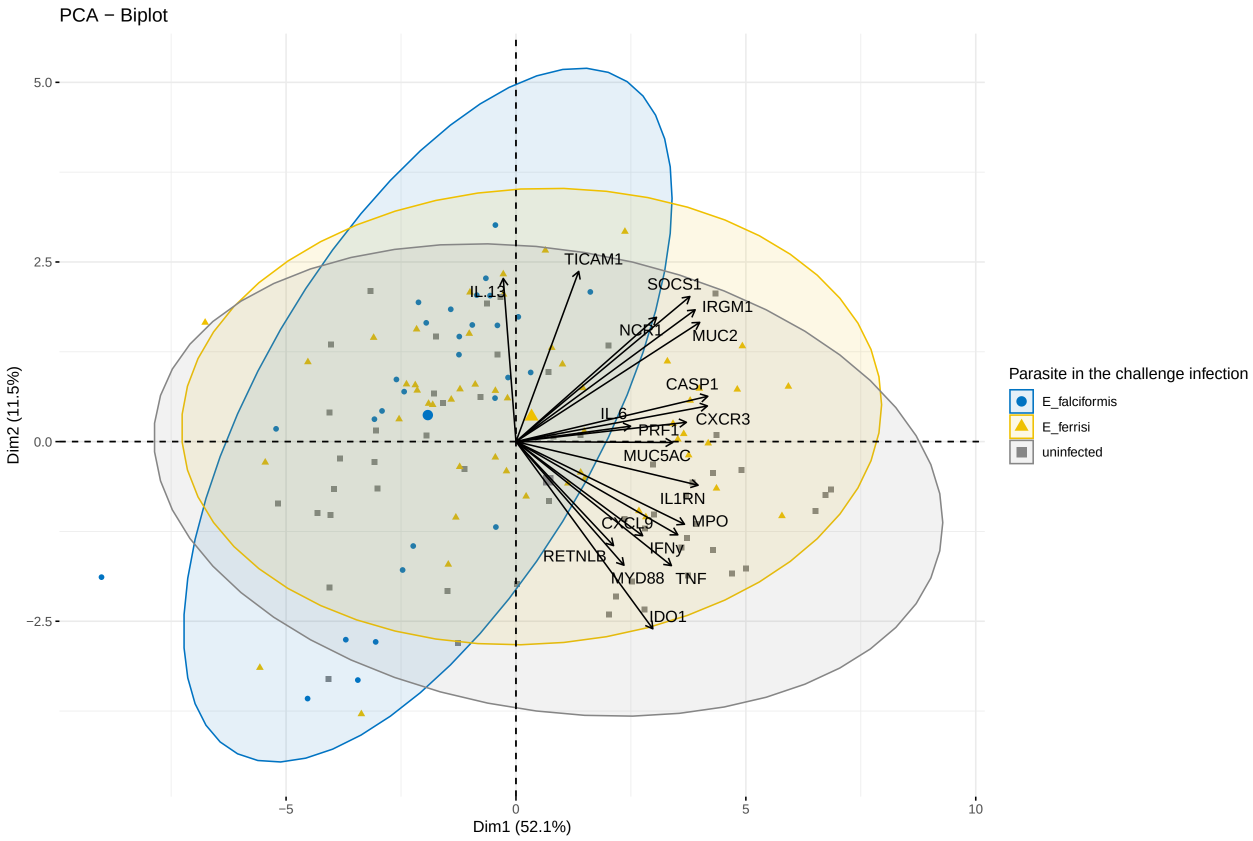
C1 and C2 are the contributions of the variable on PC1 and PC2, respectively Eig1 and Eig2 are the eigenvalues of PC1 and PC2, respectively. Recall that eigenvalues measure the amount of variation retained by each PC. In this case, the expected average contribution (cutoff) is calculated as follow: As 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



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:

```
##
## 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
```

```
##
## 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
```

```
##
```

```
## Call:
## lm(formula = WL_max ~ pc1 + pc2 + Parasite_challenge + hybrid_status,
##     data = lab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.228  -4.572   0.441   4.382  17.116
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -14.2797     1.4298  -9.987  < 2e-16 ***
## pc1                                0.3222     0.2233   1.443  0.15166
## pc2                               -1.5587     0.4743  -3.287  0.00132 **
## Parasite_challengeE_ferrisi        6.6203     1.4849   4.458 1.81e-05 ***
## Parasite_challengeuninfected       7.9023     1.5395   5.133 1.06e-06 ***
## hybrid_statusF0 M. m. musculus    -3.4993     1.4431  -2.425  0.01674 *
## hybrid_statusF1 hybrid             1.1644     2.0037   0.581  0.56219
## hybrid_statusF1 M. m. domesticus  -2.9842     2.6556  -1.124  0.26329
## hybrid_statusF1 M. m. musculus     1.4476     3.0530   0.474  0.63622
## hybrid_statusother                -0.8279     1.7624  -0.470  0.63934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.107 on 125 degrees of freedom
## Multiple R-squared:  0.4455, Adjusted R-squared:  0.4055
## F-statistic: 11.16 on 9 and 125 DF,  p-value: 1.175e-12
```

```
## [1] 883.2645
```

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)

```
weight_lm3 <- lm(WL_max ~ pc1 + pc2 + hybrid_status, data = lab)
weight_no_pc1 <- lm(WL_max ~ pc2 + hybrid_status, data = lab)
weight_no_pc2 <- lm(WL_max ~ pc1  + hybrid_status, data = lab)
weight_no_hybrid <- lm(WL_max ~ pc1 + pc2, data = lab)
lrtest(weight_lm3, weight_no_pc1)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc1 + pc2 + hybrid_status
## Model 2: WL_max ~ pc2 + hybrid_status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -444.48
## 2   8 -447.20 -1 5.4503    0.01956 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(weight_lm3, weight_no_pc2)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc1 + pc2 + hybrid_status
## Model 2: WL_max ~ pc1 + hybrid_status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -444.48
## 2   8 -453.06 -1 17.159  3.438e-05 ***
## ---
## 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 + hybrid_status
## Model 2: WL_max ~ pc1 + hybrid_status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -447.20
## 2   8 -453.06  0 11.709  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(weight_lm3, weight_no_hybrid)
```

```
## Likelihood ratio test
##
## Model 1: WL_max ~ pc1 + pc2 + hybrid_status
## Model 2: WL_max ~ pc1 + pc2
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -444.48
## 2   4 -448.63 -5 8.3132     0.1398
```

```
##
## 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
```

```
##
```

```
## Call:
## lm(formula = WL_max ~ pc1 + pc2 + infection_history, data = lab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0095  -3.8847   0.1607   4.4701  18.0112
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -14.8914     1.7659  -8.433 7.29e-14
## pc1                                       0.4849     0.2031   2.388  0.01844
## pc2                                      -2.0536     0.3912  -5.249 6.43e-07
## infection_historyfalciformis_ferrisi      6.0667     2.1863   2.775  0.00638
## infection_historyfalciformis_uninfected   5.1523     2.2551   2.285  0.02403
## infection_historyferrisi_falciformis     -1.4411     2.5744  -0.560  0.57664
## infection_historyferrisi_ferrisi          6.9461     2.2641   3.068  0.00265
## infection_historyferrisi_uninfected       6.9622     2.3587   2.952  0.00378
## infection_historyuninfected              12.9012     2.7488   4.693 6.99e-06
## infection_historyuninfected_falciformis   0.6748     3.0346   0.222  0.82440
## infection_historyuninfected_ferrisi       4.6608     2.7644   1.686  0.09431
##
## (Intercept)                             ***
## pc1                                     *
## pc2                                     ***
## infection_historyfalciformis_ferrisi    **
## infection_historyfalciformis_uninfected *
## infection_historyferrisi_falciformis
## infection_historyferrisi_ferrisi        **
## infection_historyferrisi_uninfected     **
## infection_historyuninfected             ***
## infection_historyuninfected_falciformis
## infection_historyuninfected_ferrisi     .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.107 on 124 degrees of freedom
## Multiple R-squared:  0.4499, Adjusted R-squared:  0.4055
## F-statistic: 10.14 on 10 and 124 DF,  p-value: 2.583e-12

## [1] 884.1813

##
## 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
```

```
##                     df      AIC
## weight_lm            9 906.9537
## weight_lm_exp_only  4 905.2669
```

repeating the heatmap on the now imputed data

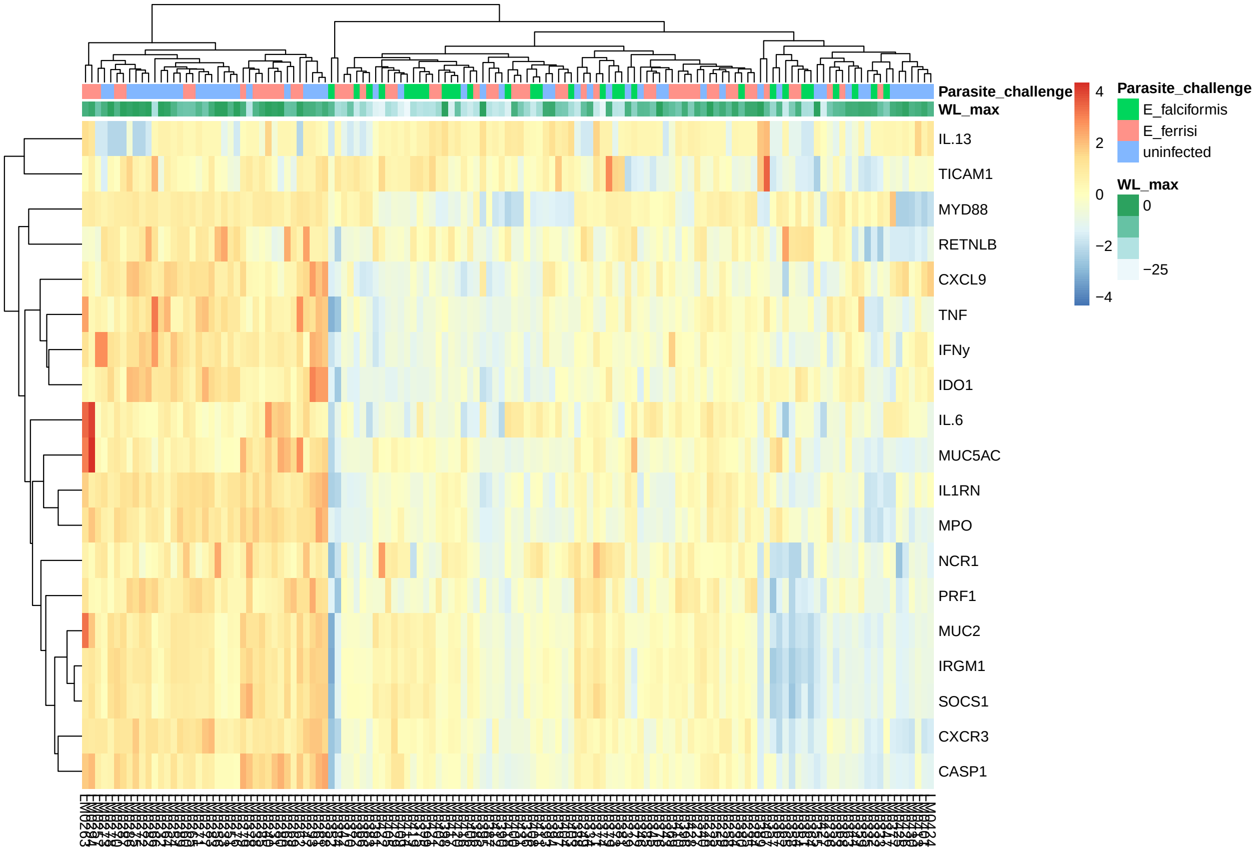
```
 # turn the data frame into a matrix and transpose it. We want to have each cell
 # type as a row name
 gene <- t(as.matrix(gene))

 # 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)]


#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:


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
Parasite_challenge
E_falciformis
E_ferrisi
uninfected
WL_max