facs_gene_analysis

Fay

2022-05-19

FACS - genes Imputation and pca

Load the data sets

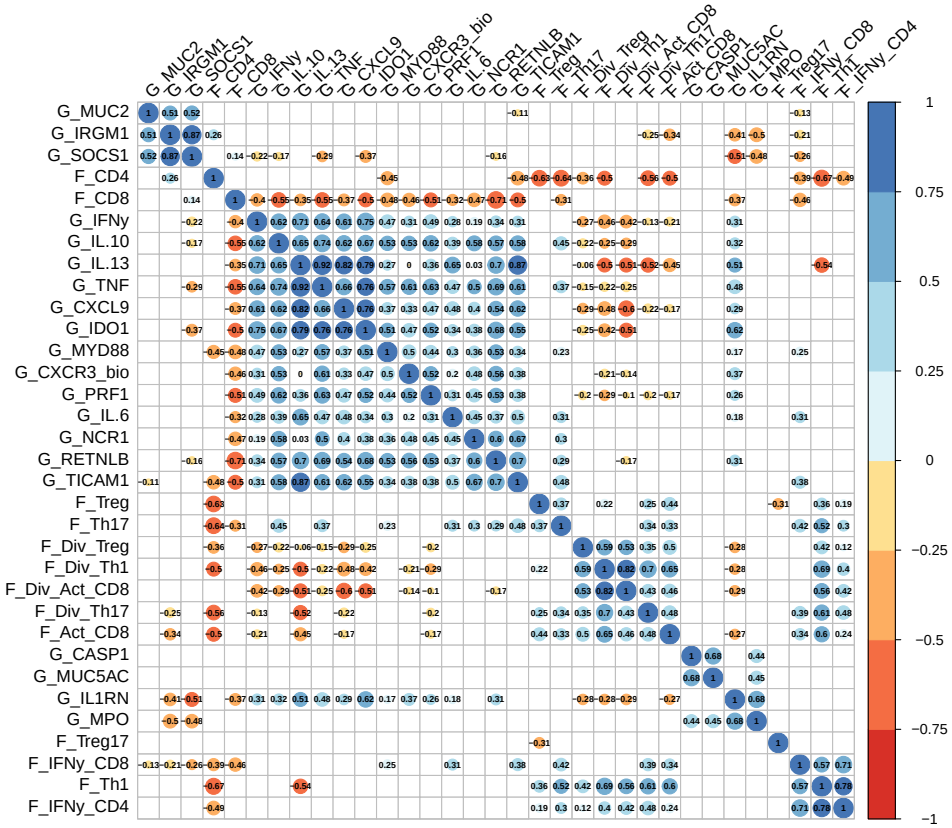
```
facs <- read.csv("output_data/facs/data_products/FACS_clean.csv")
gene <- read.csv("output_data/gene_expression/data_products/clean_gene_expression.csv")
```

1. Start by combining the data sets

```
## Adding prefixes to the columns of each data frame and joining

#Adding the suffix G to the genes
colnames(gene) <- paste("G_", colnames(gene), sep = "")

gene <- gene %>% rename(EH_ID = G_EH_ID,
                                  primary_infection = G_primary_infection,
                                  challenge_infection = G_challenge_infection,
                                  infection_history = G_infection_history,
                                  mouse_strain = G_mouse_strain,
                                  max_WL = G_max_WL,
                                  delta = G_delta,
                                  Parasite_challenge =  G_Parasite_challenge,
                                  hybrid_status = G_hybrid_status)

#Adding the suffix f to the facs data
colnames(facs) <- paste("F_", colnames(facs), sep = "")

facs <- facs %>% rename(EH_ID = F_EH_ID,
                        infection_history = F_infection_history,
                        max_WL = F_max_WL,
                        Parasite_challenge =  F_Parasite_challenge,
                        hybrid_status = F_hybrid_status,
                        delta = F_delta)

immune_data <- gene %>% full_join(facs, by = intersect(colnames(gene), colnames(facs)))

immune_data <- unique(immune_data)
```

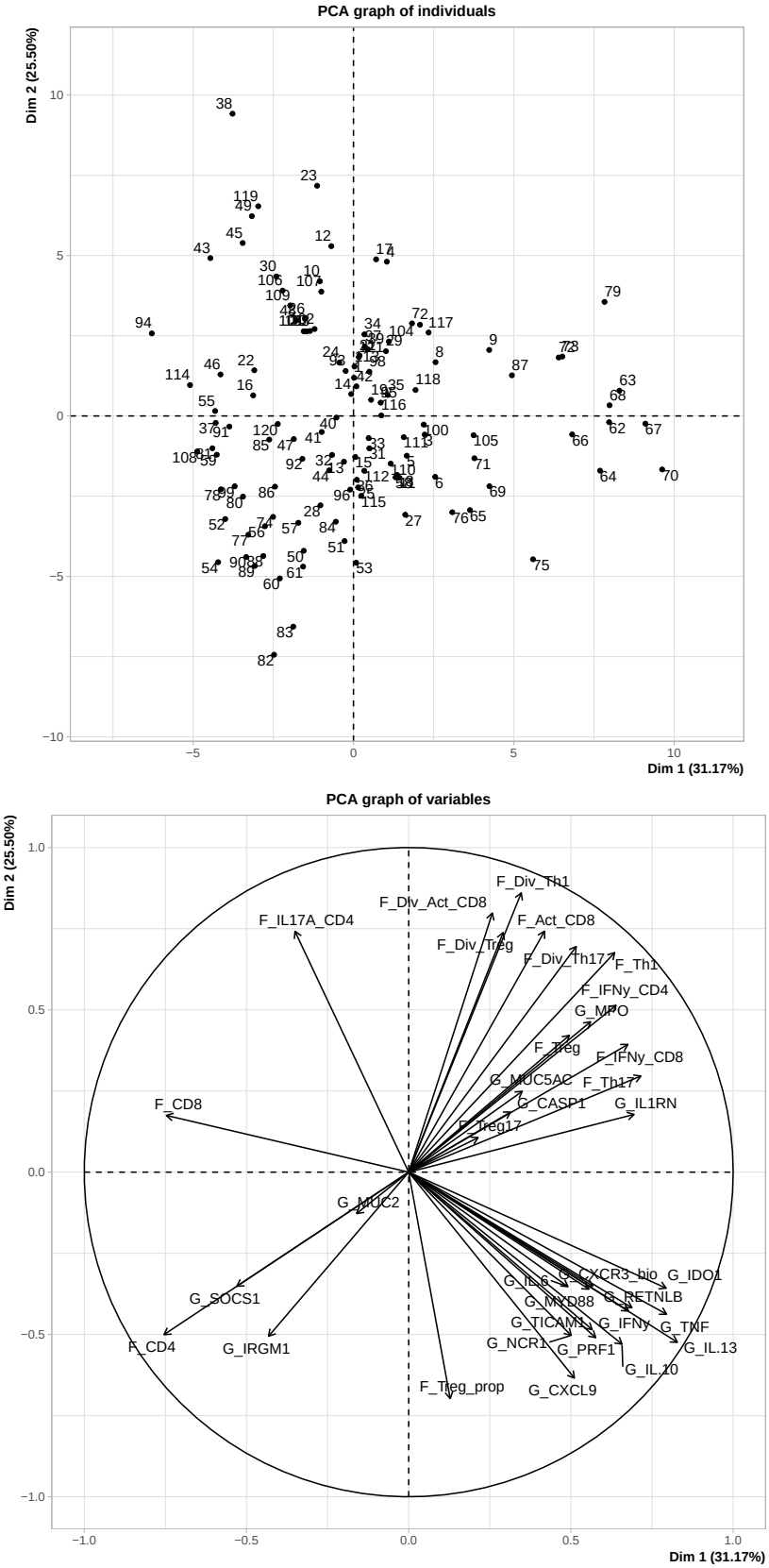
Now we go on to see the correlations between our data

```
corrplot(immune_correlation,
         method = "circle",  #method of the plot, "color" would show colour gradient
         tl.col = "black", tl.srt = 45, #colour of labels and rotation
         col = brewer.pal(n = 8, name ="RdYlBu"), #colour of matrix
         order="hclust", #hclust reordering
         p.mat = p.mat, sig.level = 0.01, insig = "blank",
         addCoef.col = 'black',
         number.cex = 0.5)
```



We will now continue by using an iterative pca to impute missing data A. Initialization: impute using the mean B. Step lampda: # a. do pca on imputed data table S dimensions retained # b. missing data imputed using pca # c. means (and standard deviations) updated C. Iterate the estimation and imputation steps (until convergence) (convergence: the act of converging and especially moving toward union or uniformity)

Overfitting is a common problem due to believing too much in links between variables. –> regularized iterative PCA (This version is what is being implented in missMDA) This is a way of taking less risk when imputing the missing data. The algorithm estimates the missing data values with values that have no influence on the PCA results, i.e., no influence on the coordinates of the individals or variables.



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

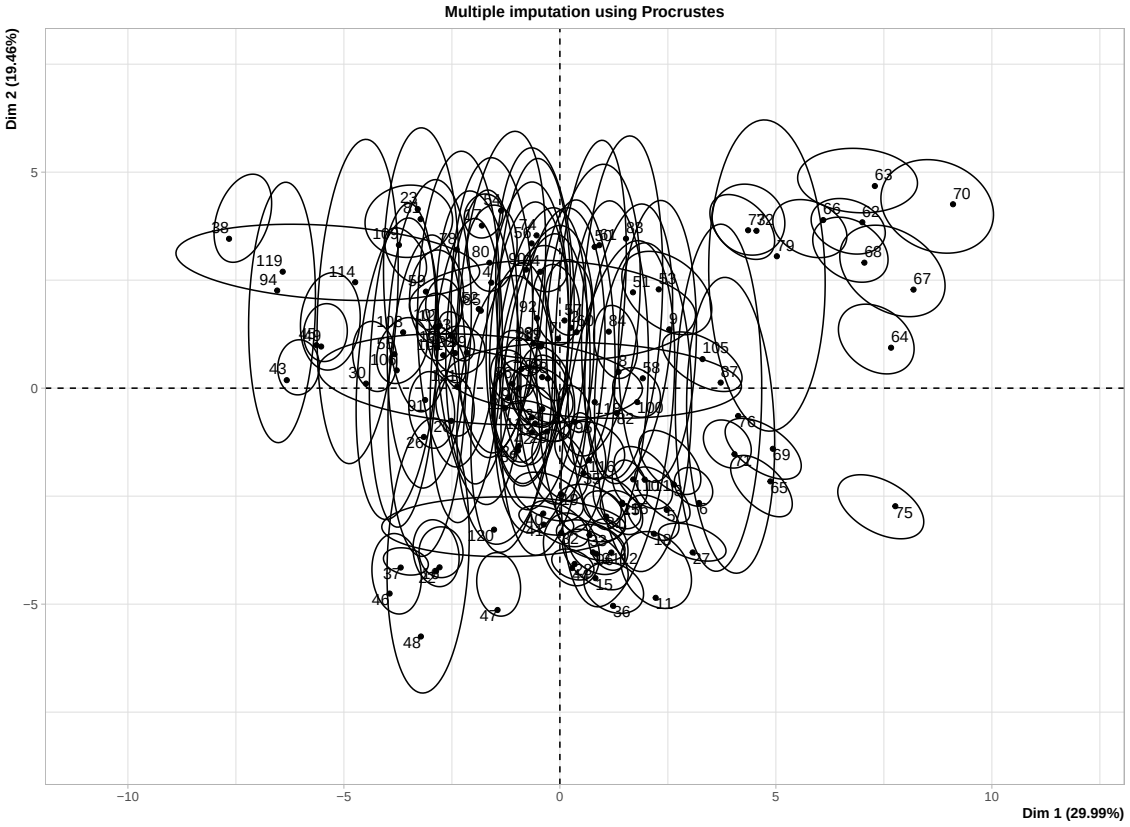
estimated.

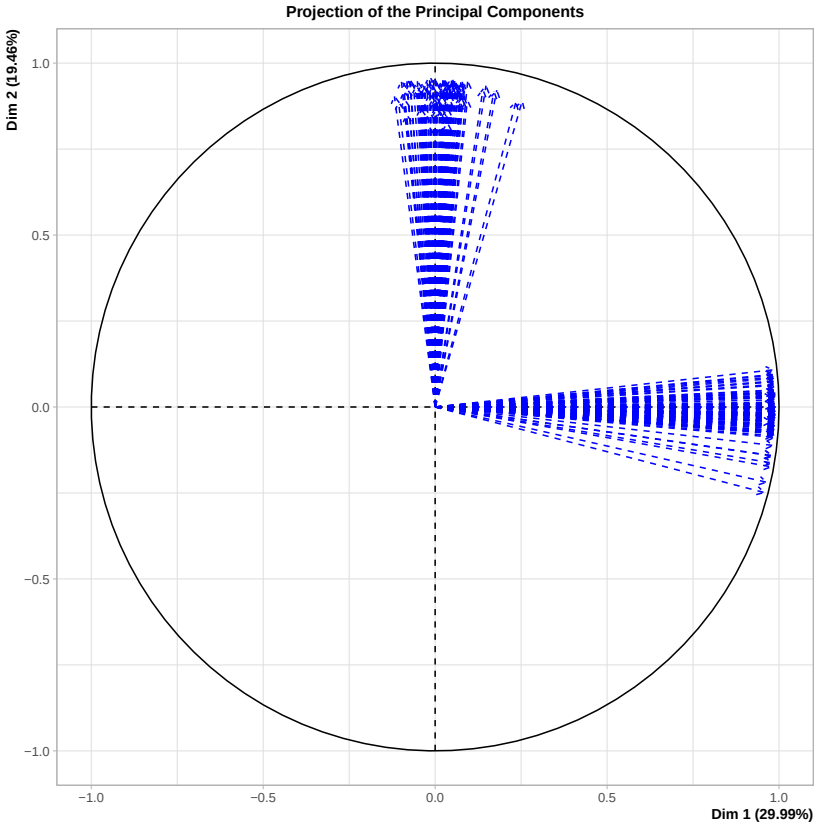
Another problem: the imputed data are, when the pca is performed considered like real observations. But they are estimations!!

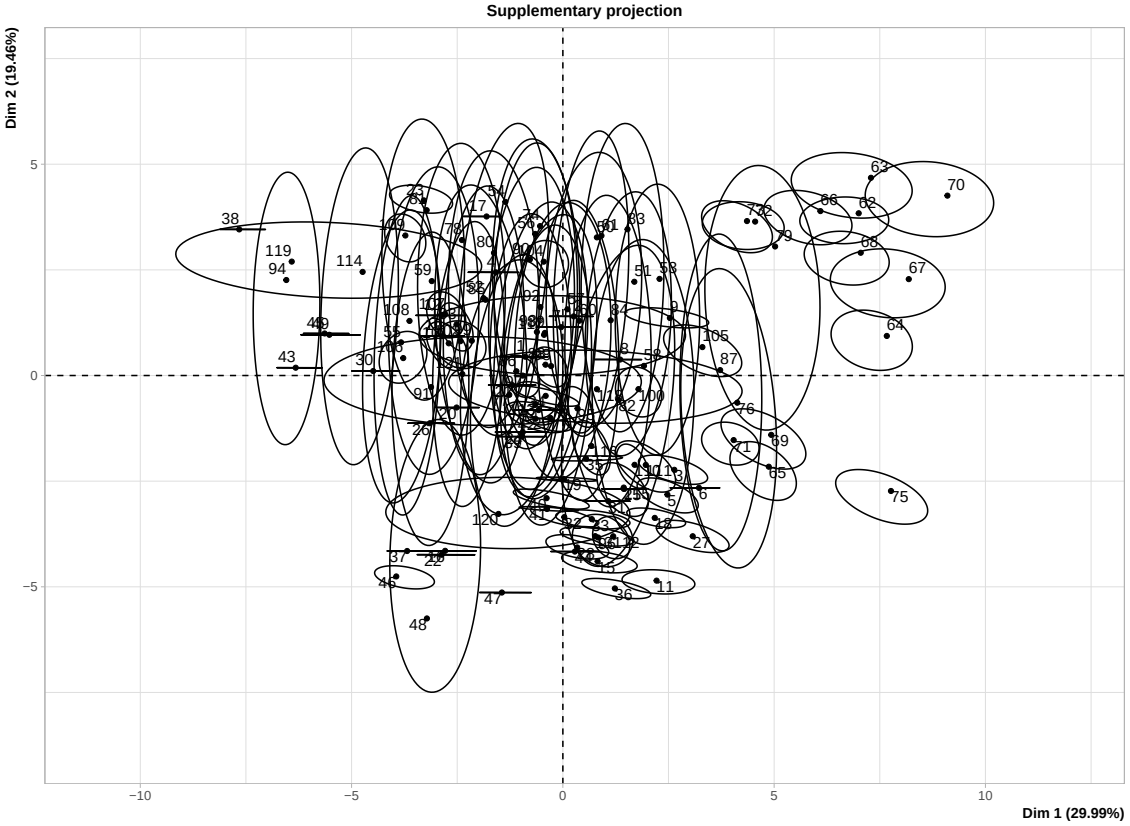
Visualizing uncertainty due to issing data:

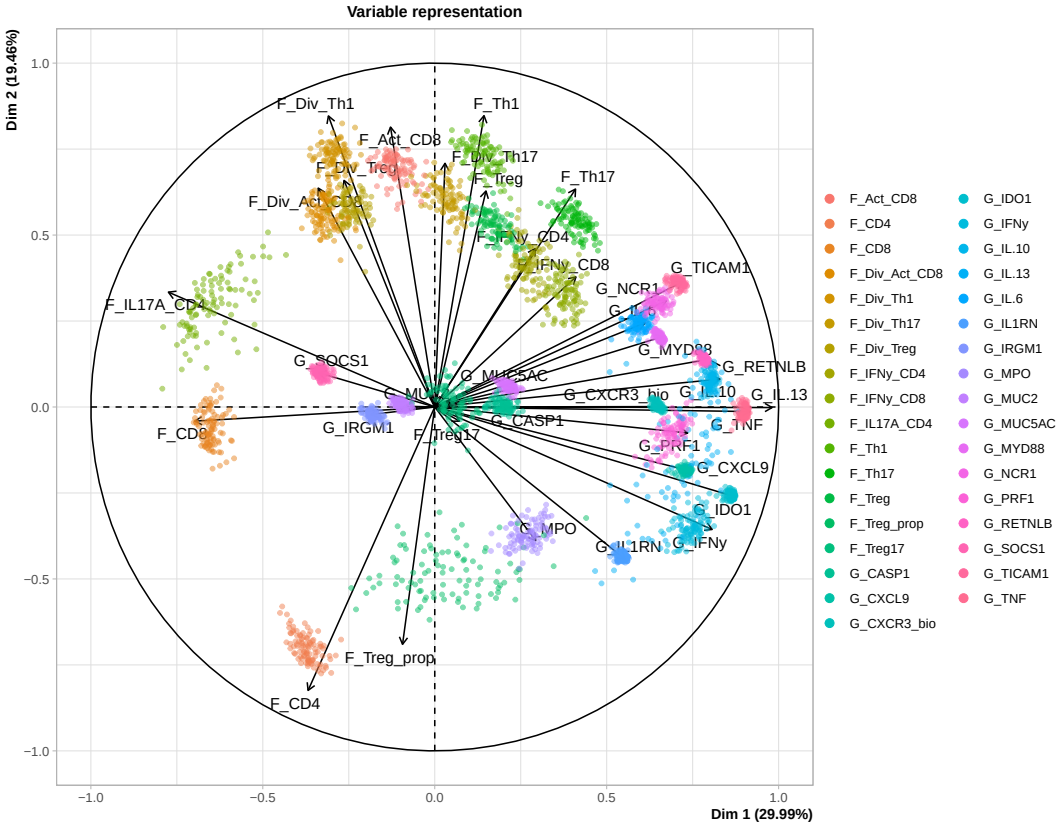
–> mulrimple imputation: generate several plausible values for each missing data point

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

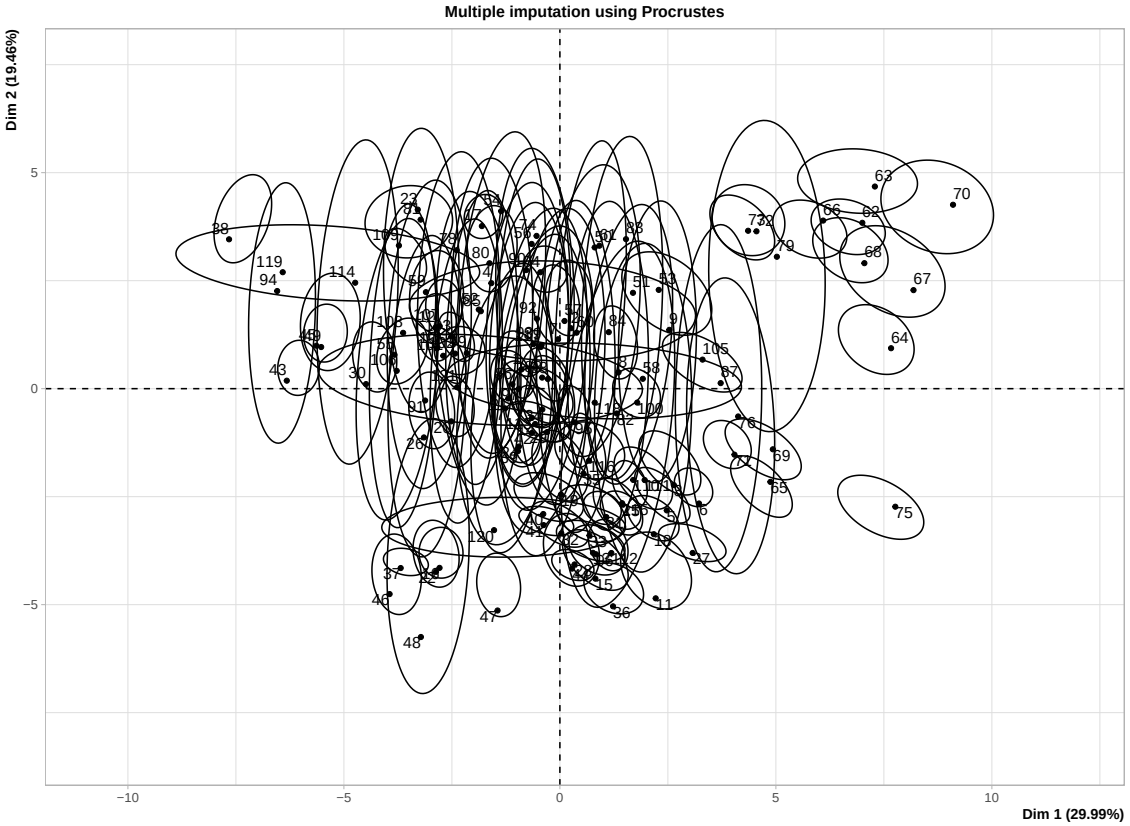



Projection of the Principal Components
Dim 2 (19.46%)
Dim 1 (29.99%)
1.0
0.5
0.0
−0.5
−1.0

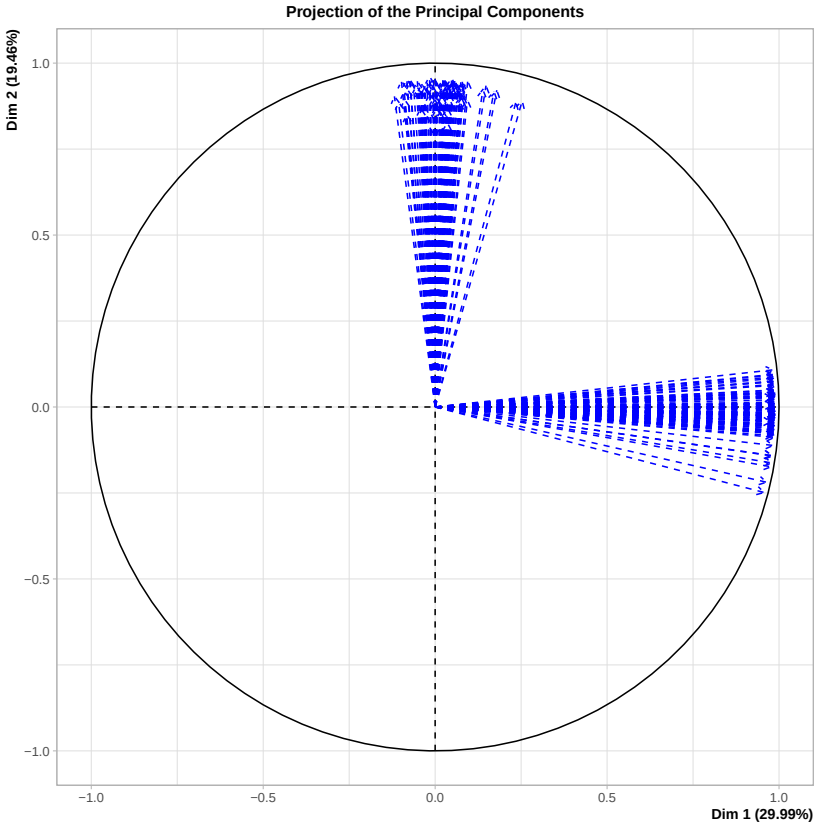

Supplementary projection
Dim 2 (19.46%)
Dim 1 (29.99%)
5
0
−5
−10
10



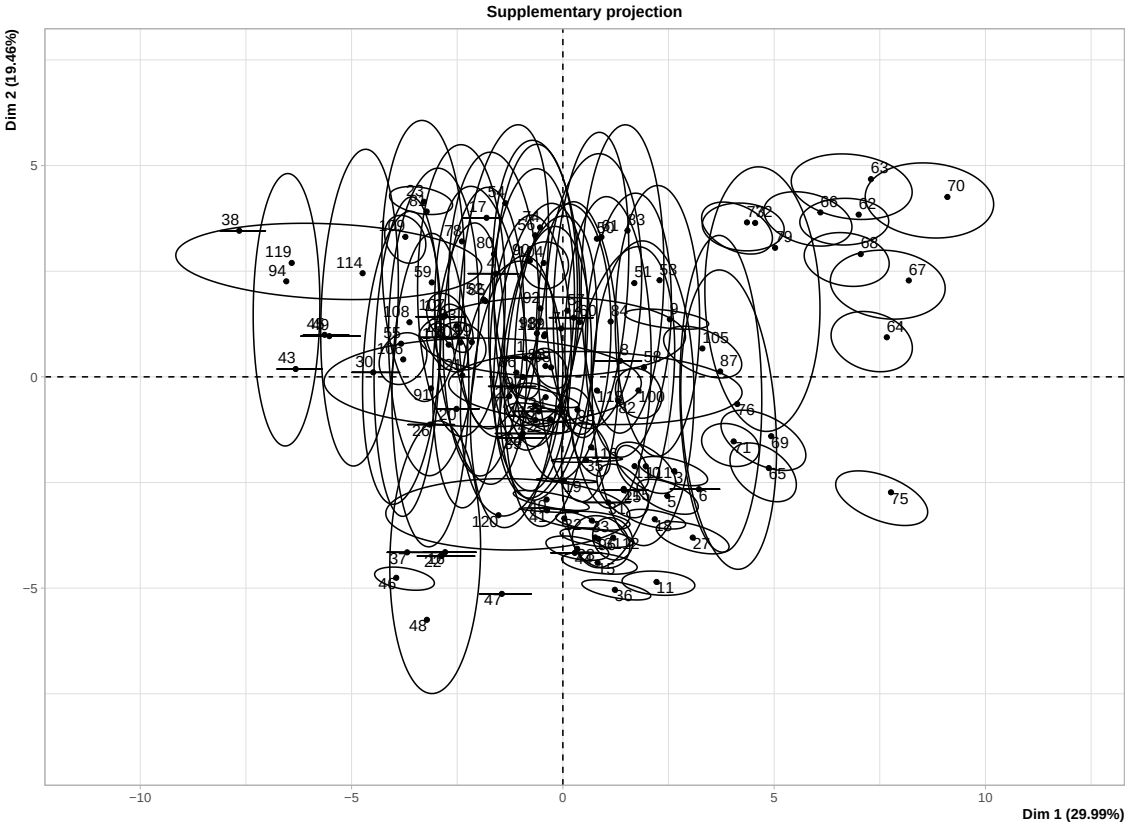
$PlotIndProc


Multiple imputation using Procrustes
Dim 2 (19.46%)
Dim 1 (29.99%)
5
0
−5
−10
−5
0
5
10

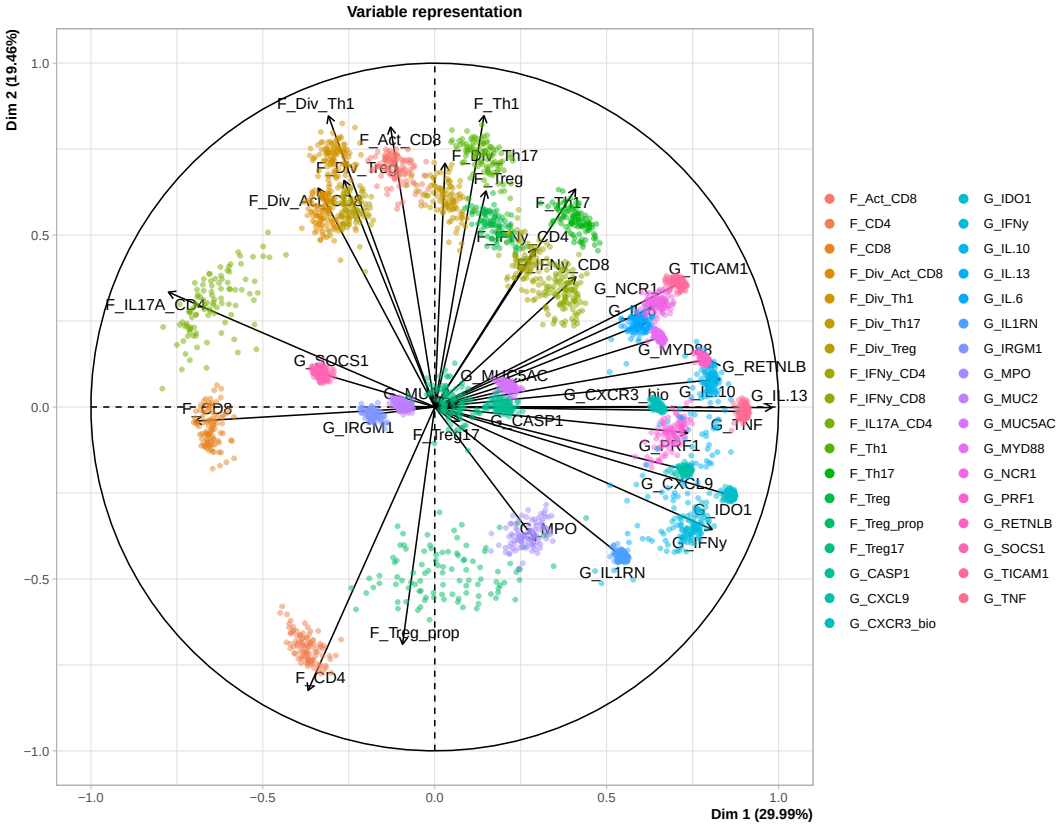
```
##
## $PlotDim
```




```
##
## $PlotIndSupp
```




```
##
## $PlotVar
```



Individuals lying on the axis have no missing data, but individuals that far away have many missing data. big ellipse = big uncertainty tight elipse (line) = low uncertainty

Variable representation: Poins tight together)look like one) - have no missing variables –> low uncertainty Points spread – > higher variability – > higher uncertainty

High uncertainty–> we should interpret the result with care

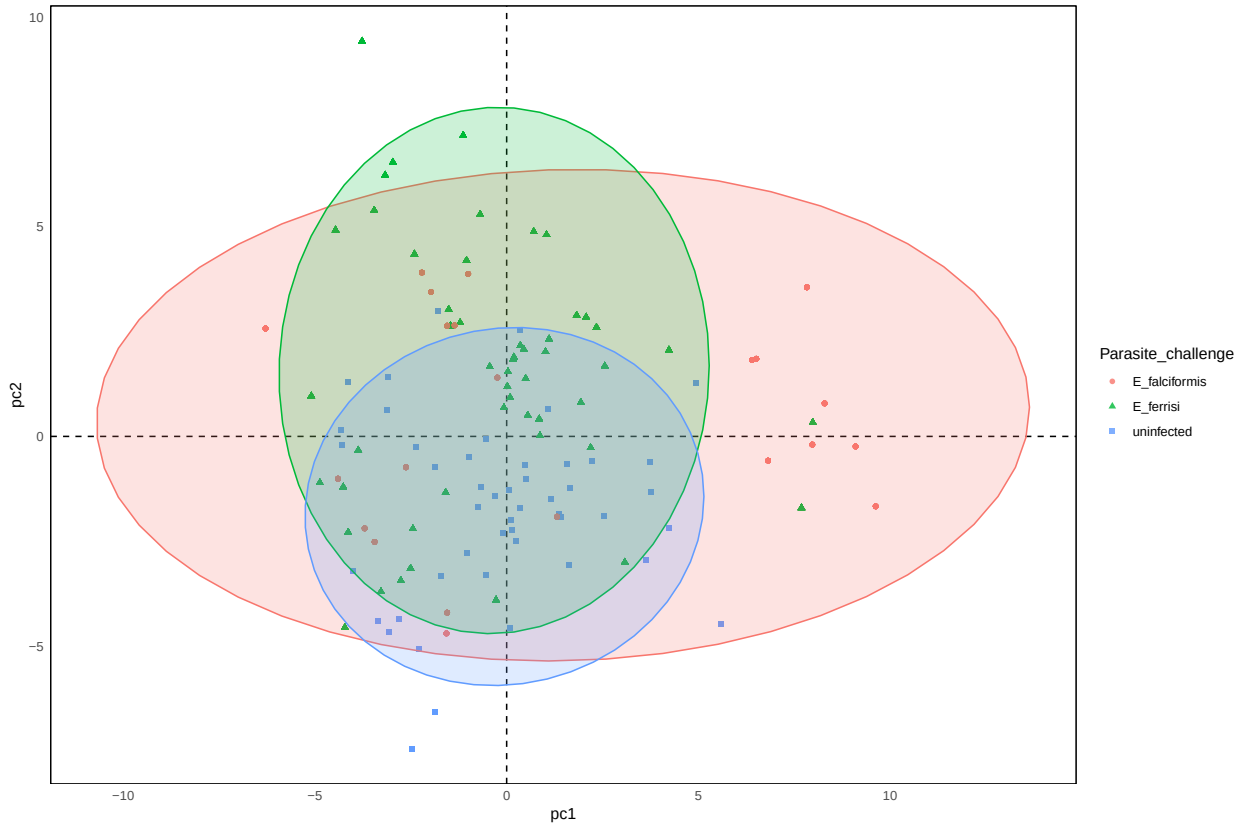
The individuals with many missing data values make the axes move, and thus the positions of all individuals

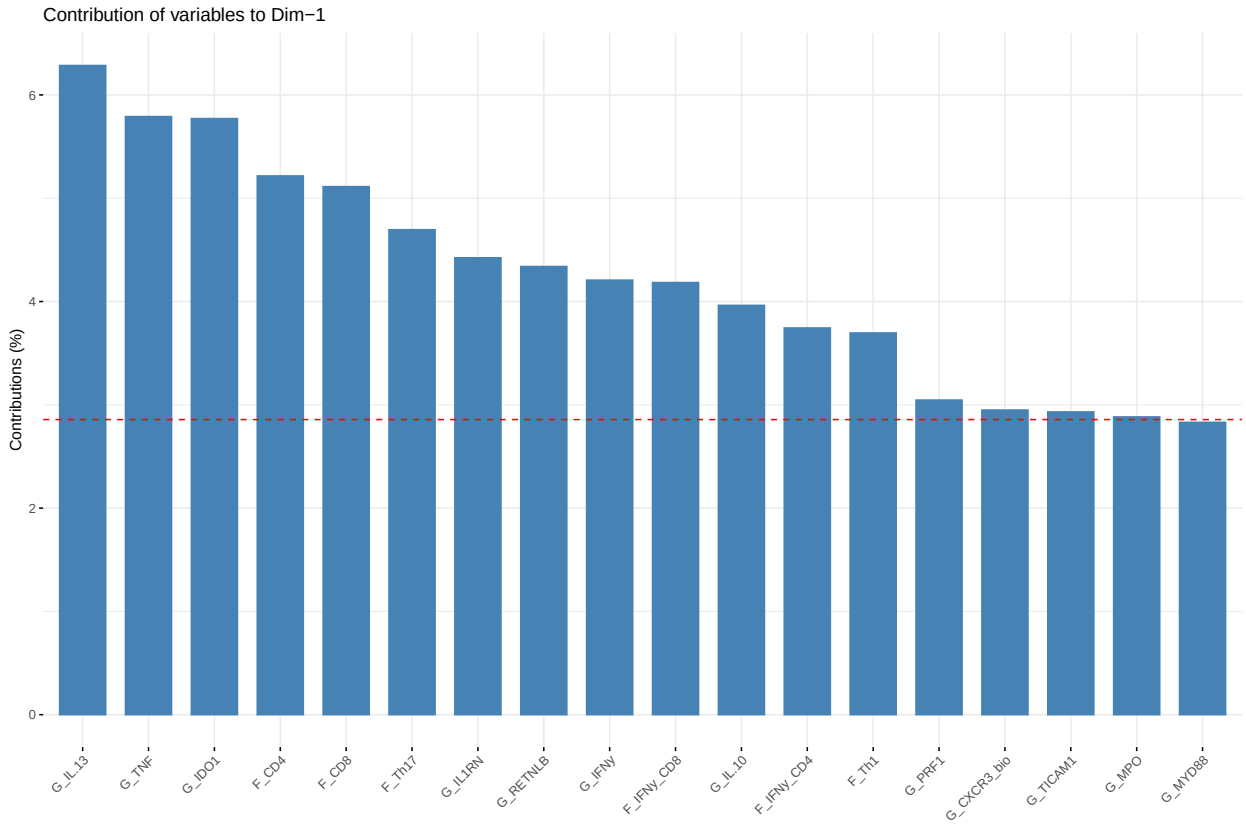
Therefore in the last plots every individual is getting an eclipse as they are as well influenced by the missing data of the others.

THe plot with the dimensions shows the projections of the pca dimensions of each imputed table on the pca plane obtained using the original imputed data table

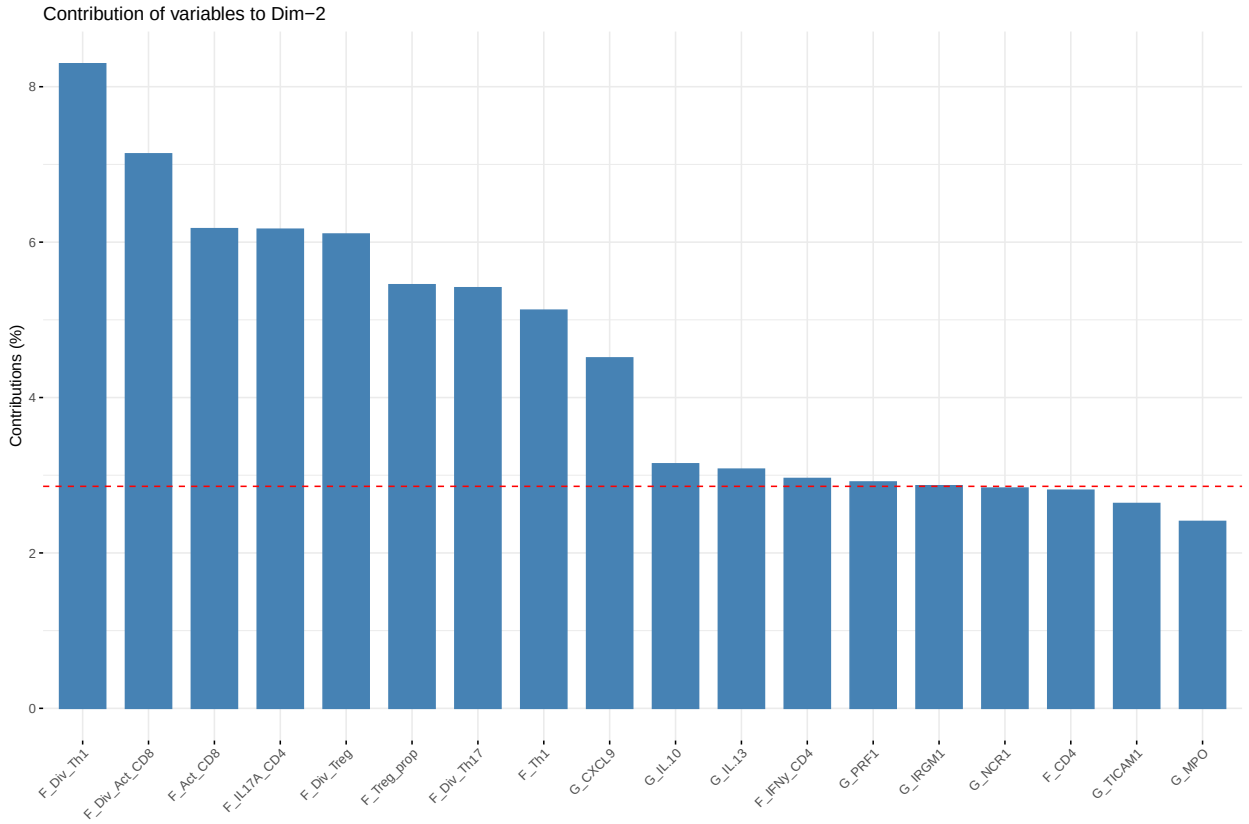
As all of the arrows are close to either the first or second axes, this means that the axes are stable with respect to the set of imputed tables –> we don't have evidence of instability here.

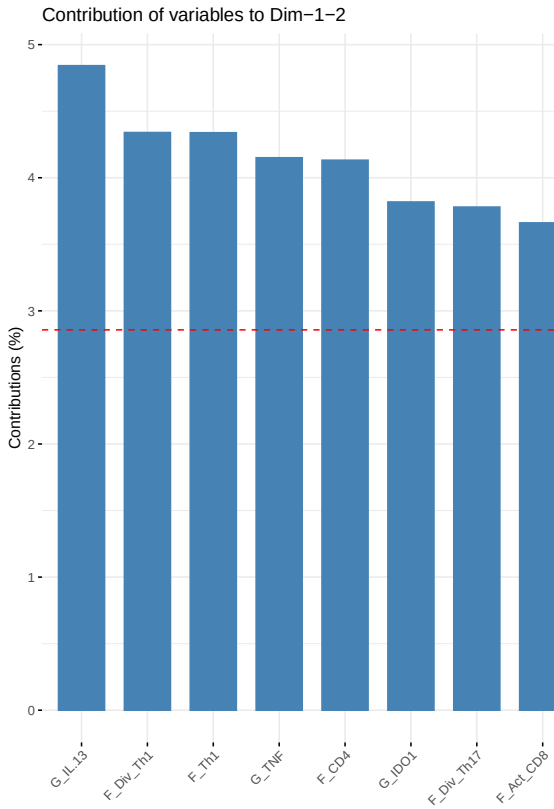
```
imputed_immune %>%
  pivot_longer(cols = c(13:47), names_to = "Immune_parameter", values_to = "measurement") %>%
  ggplot(aes(x = pc1, y = pc2, color = Parasite_challenge, shape = Parasite_challenge)) +
  geom_hline(yintercept = 0, lty = 2) +
  geom_vline(xintercept = 0, lty = 2) +
  geom_point(alpha = 0.8) +
  stat_ellipse(geom="polygon", aes(fill = Parasite_challenge), alpha = 0.2, show.legend = FALSE,
               level = 0.95) +
  theme_minimal() +
  theme(panel.grid = element_blank(), panel.border = element_rect(fill= "transparent"))
```


pc2
pc1
10
5
0
−5
−10
−5
0
5
10
Parasite_challenge
E_falciformis
E_ferrisi
uninfected


Contribution of variables to Dim−1
Contributions (%)
6
4
2
0
G_IL13
G_TNF
G_IDO1
F_CD4
F_CD8
F_Th17
G_IL1RN
G_RETNLB
G_IFNγ
F_IFNγ_CD8
G_IL10
F_IFNγ_CD4
F_Th1
G_PRF1
G_CXCR3_bio
G_TICAM1
G_MPO
G_MYD88

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





The total contribution to PC1 and PC2 is obtained with the following R code:

Linear models:

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + Parasite_challenge, data = imputed_immune)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4074  -3.0426   0.1079   3.4467  14.5417
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   85.0389     1.1506  73.907  < 2e-16 ***
## pc1                            0.2600     0.1490   1.746   0.0835 .
## pc2                            0.3444     0.1841   1.871   0.0638 .
## Parasite_challengeE_ferrisi    6.3924     1.3932   4.588 1.14e-05 ***
## Parasite_challengeuninfected  11.6701     1.4368   8.122 5.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.29 on 116 degrees of freedom
## Multiple R-squared:  0.3667, Adjusted R-squared:  0.3449
## F-statistic: 16.79 on 4 and 116 DF,  p-value: 6.934e-11

## [1] 753.4251

##
```

```
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + Parasite_challenge + hybrid_status,
##     data = imputed_immune)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2450  -3.1938   0.7664   3.5383  14.3204
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      85.35475    1.39517  61.179  < 2e-16 ***
## pc1                               0.25489    0.17185   1.483    0.141
## pc2                               0.27613    0.22418   1.232    0.221
## Parasite_challengeE_ferrisi       6.07832    1.44090   4.218 5.04e-05 ***
## Parasite_challengeuninfected     10.99314    1.57720   6.970 2.39e-10 ***
## hybrid_statusF0 M. m. musculus   -1.13546    1.48570  -0.764    0.446
## hybrid_statusF1 hybrid            2.17931    1.61285   1.351    0.179
## hybrid_statusF1 M. m. domesticus -1.83057    2.07730  -0.881    0.380
## hybrid_statusF1 M. m. musculus    2.35702    2.35212   1.002    0.318
## hybrid_statusother               -0.05226    1.48003  -0.035    0.972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.266 on 111 degrees of freedom
## Multiple R-squared:  0.3996, Adjusted R-squared:  0.3509
## F-statistic: 8.208 on 9 and 111 DF,  p-value: 2.824e-09

## [1] 756.9754

##
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + hybrid_status, data = imputed_immune)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1398  -3.4005   0.9191   4.8726  10.6773
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      92.04227    1.00974  91.155  < 2e-16 ***
## pc1                               0.26621    0.20118   1.323  0.18844
## pc2                              -0.40949    0.22097  -1.853  0.06646 .
## hybrid_statusF0 M. m. musculus   -0.52083    1.75427  -0.297  0.76709
## hybrid_statusF1 hybrid            4.91073    1.85958   2.641  0.00944 **
## hybrid_statusF1 M. m. domesticus -0.07881    2.44470  -0.032  0.97434
## hybrid_statusF1 M. m. musculus    5.65349    2.73604   2.066  0.04109 *
## hybrid_statusother               -2.73685    1.69947  -1.610  0.11010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.262 on 113 degrees of freedom
## Multiple R-squared:  0.1356, Adjusted R-squared:  0.08209
## F-statistic: 2.533 on 7 and 113 DF,  p-value: 0.01853
```

```
## [1] 797.0621
```

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + infection_history, data = imputed_immune)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3211  -2.9505  -0.1901   3.0973  14.1793
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             89.5205     1.9225  46.564  < 2e-16
## pc1                                      0.1766     0.1439   1.228 0.222170
## pc2                                      0.3565     0.1826   1.952 0.053440
## infection_historyfalciformis_ferrisi     1.9627     2.2774   0.862 0.390661
## infection_historyfalciformis_uninfected  8.0188     2.3029   3.482 0.000715
## infection_historyferrisi_falciformis    -7.8996     2.5523  -3.095 0.002496
## infection_historyferrisi_ferrisi         3.5911     2.2650   1.585 0.115732
## infection_historyferrisi_uninfected      6.4421     2.1645   2.976 0.003588
## infection_historyuninfected              8.1959     2.6371   3.108 0.002398
## infection_historyuninfected_falciformis -4.1565     2.7876  -1.491 0.138804
## infection_historyuninfected_ferrisi     -2.9895     2.5725  -1.162 0.247723
##
## (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: 4.948 on 110 degrees of freedom
## Multiple R-squared:  0.4746, Adjusted R-squared:  0.4268
## F-statistic: 9.937 on 10 and 110 DF,  p-value: 1.059e-11
```

```
## [1] 742.8219
```

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2, data = imputed_immune)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.397  -3.143   1.903   4.982   8.477
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 92.36267    0.59806 154.438   <2e-16 ***
## pc1          0.07067    0.18107   0.390    0.697
## pc2         -0.11052    0.20019  -0.552    0.582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.579 on 118 degrees of freedom
## Multiple R-squared:  0.003859,   Adjusted R-squared:  -0.01302
## F-statistic: 0.2286 on 2 and 118 DF,  p-value: 0.796
```

```
##                     df      AIC
## weight_lm            6 753.4251
## weight_lm_exp_only   4 804.2317
```

repeating the heatmap on the now imputed data

Heatmap on imputed combined data:

```
#plot the heatmap

heatmap_data %>%
  pheatmap(annotation_col = annotation_df, scale = "row")
```