Gene_expression_analysis

Fay

2022-05-18

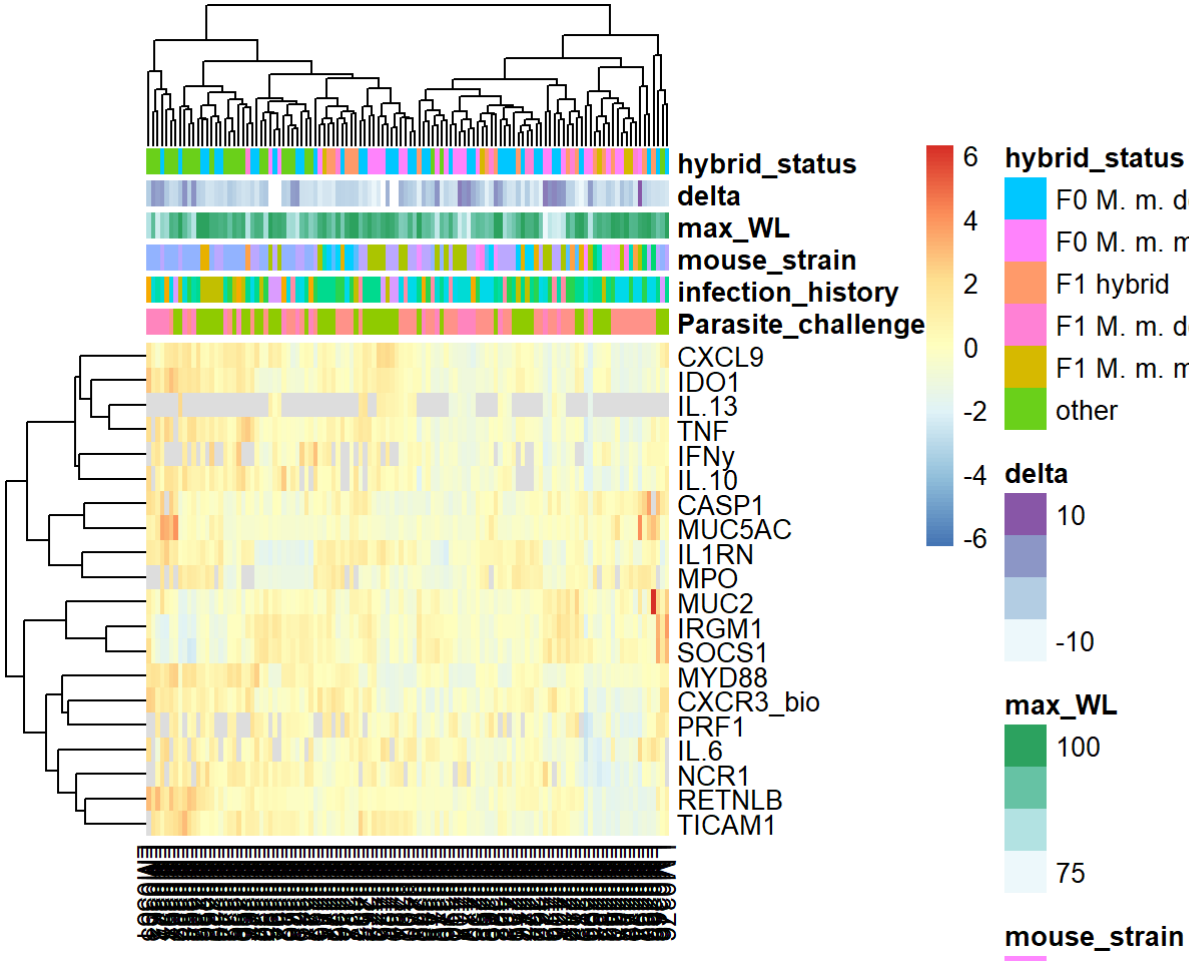
GAPDH HKG
B-actin HKG
Ppia HKG
Ppip HKG
CDC42 HKG susceptible to DNA contamination

Relm-b mucosal defense factor (goblet cells)
Muc2 the major secretory mucin within the gastrointestinal tract
TFF3 mucosal defense factor (goblet cells)
Muc5ac similar to MUC2, produced by surface goblet cells
NKp46 NK marker
F4/80 macropahge marker (distinguish by immune response trend)
Mpo myeloperoxidase in Neutrophils
MyD88 TLR protein, NF-kB IRAK protein, inflammation marker by TLR MyD88-Dependent Pathway caspase-1 inflammasome marker (IL-1b and IL-18 production)
IL-1Ra natuaral IL-1b antagonist for infection control (if not increase in Tregs is seen)
CXCL9, immune cell migration marker + Th1 activator (confirm FACS)
CXCR3, CXCL9 and CXCL11 receptor
IL-6 TNF inhibitor,
IL-12ra T-cell marker Th1
IFN-y compare with IFN-y producing cells and IFN-y ELISAs, should correlate wit PRF1, NKp46 and F4/80. One of these cell types just have to be doing the job!
IRG6A autonomous cell defense (opsonization)
TNF-a upregulated in eimeria but not well explained. Could be present and driving infection where IFN-y isn't
IL-17 in case IFN-y isn't coming up but pathogenicity is
TRIF Type I IFN production TRIF Dependent Pathway
Socs1 JAK/STAT signaling pathway, proinflammatory regulating + T-cell differentiation, could explain severity
IDO1 DC, monocyte and MC protein regulating T-cell activity
Prf1 perforin, should be dominant in primary infections, but must be correlated between T-cell and NK cell expresions
CD56 CD56bright = more cytokine producing NKs, CD56dim = more direct cytotoxic killing
IL-4
IL-13
IL-10

1. Gene expression in the laboratory infections - Heatmap

```
##
## FALSE  TRUE
##    17     3
```

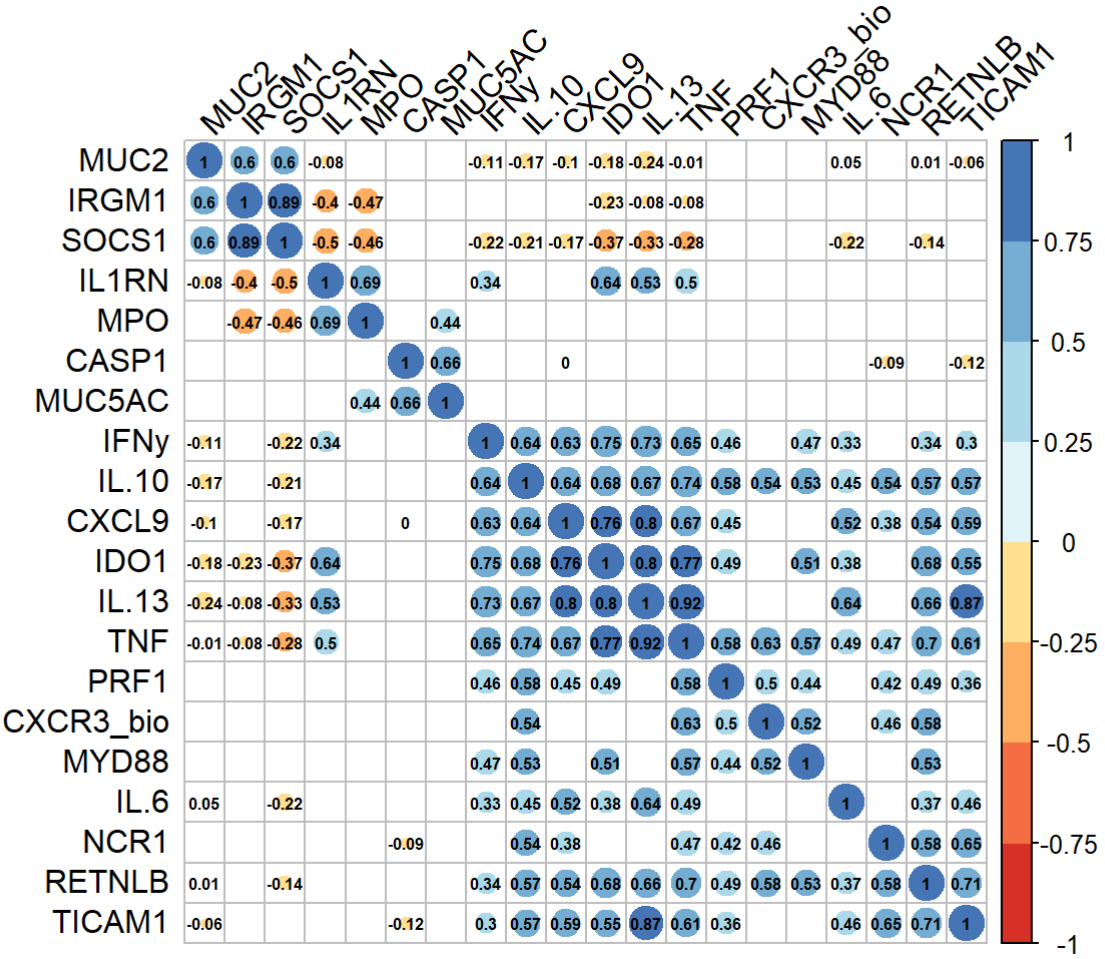
Heatmap on gene expression data:



2. Correlations between the genes

Corrplot of correlations

Here is a corrplot of the correlations between the genes. I am using the non-normalized genes



Group 1:

Muc2 the major secretory mucin within the gastrointestinal tract Mucin 2 is particularly prominent in the gut where it is secreted from goblet cells in the epithelial lining into the lumen of the large intestine. There, mucin 2, along with small amounts of related-mucin proteins, polymerizes into a gel of which 80% by weight is oligosaccharide side-chains that are added as post-translational modifications to the mucin proteins. This gel provides an insoluble mucous barrier that serves to protect the intestinal epithelium.

IRGM: Immunity-related GTPase family M protein (IRGM), also known as interferon-inducible protein 1 (IFI1), is an enzyme that in humans is IRGM gene.[5]

IRGM is a member of the interferon-inducible GTPase family. The encoded protein may play a role in the innate immune response by regulating autophagy formation in response to intracellular pathogens.

SOCS1: Suppressor of cytokine signaling 1 is a protein. SSI family members are cytokine-inducible negative regulators of cytokine signaling. The expression of this gene can be induced by a subset of cytokines, including IL2, IL3 erythropoietin (EPO), GM-CSF, and interferon-gamma (IFN-γ). The protein encoded by this gene functions downstream of cytokine receptors, and takes part in a negative feedback loop to attenuate cytokine signaling. Knockout studies in mice suggested the role of this gene as a modulator of IFN-γ action, which is required for normal postnatal growth and survival.[8] JAK/STAT signaling pathway, proinflammatory regulating + T-cell differentiation, could explain severity

Group 2: IFNy: IFN-γ, or type II interferon, is a cytokine that is critical for innate and adaptive immunity against viral, some bacterial and protozoan infections. IFN-γ is an important activator of macrophages and inducer of major histocompatibility complex class II molecule expression. Aberrant IFN-γ expression is associated with a number of autoinflammatory and autoimmune diseases. The importance of IFN-γ in the immune system stems in part from its ability to inhibit viral replication directly, and most importantly from its immunostimulatory and immunomodulatory effects. IFN-γ is produced predominantly by natural killer cells (NK) and natural killer T cells (NKT) as part of the innate immune response, and by CD4 Th1 and CD8 cytotoxic T lymphocyte (CTL)

effector T cells once antigen-specific immunity develops[11][12] as part of the adaptive immune response. IFN-γ is also produced by non-cytotoxic innate lymphoid cells (ILC), a family of immune cells first discovered in the early 2010s.[13]

IL10: IL-10 is a cytokine with multiple, pleiotropic, effects in immunoregulation and inflammation. It downregulates the expression of Th1 cytokines, MHC class II antigens, and co-stimulatory molecules on macrophages. It also enhances B cell survival, proliferation, and antibody production. IL-10 can block NF-κB activity, and is involved in the regulation of the JAK-STAT signaling pathway. Further investigation has shown that IL-10 predominantly inhibits lipopolysaccharide (LPS) and bacterial product mediated induction of the pro-inflammatory cytokines TNFα,[24] IL-1β,[24] IL-12,[25] and IFNγ[26] secretion from Toll-Like Receptor (TLR) triggered myeloid lineage cells. IL-10 is capable of inhibiting synthesis of pro-inflammatory cytokines such as IFN-γ, IL-2, IL-3, TNFα and GM-CSF made by cells such as macrophages and Th1 T cells. It also displays a potent ability to suppress the antigen-presentation capacity of antigen presenting cells; however, it is also stimulatory towards certain T cells (Th2) and mast cells and stimulates B cell maturation and antibody production.

CXCL9: Chemokine (C-X-C motif) ligand 9 (CXCL9) is a small cytokine belonging to the CXC chemokine family that is also known as monokine induced by gamma interferon (MIG). The CXCL9 is one of the chemokine which plays role to induce chemotaxis, promote differentiation and multiplication of leukocytes, and cause tissue extravasation.[3]

The CXCL9/CXCR3 receptor regulates immune cell migration, differentiation, and activation. Immune reactivity occurs through recruitment of immune cells, such as cytotoxic lymphocytes (CTLs), natural killer (NK) cells, NKT cells, and macrophages. Th1 polarization also activates the immune cells in response to IFN-γ.[4] Tumor-infiltrating lymphocytes are a key for clinical outcomes and prediction of the response to checkpoint inhibitors.[5] In vivo studies suggest the axis plays a tumorigenic role by increasing tumor proliferation and metastasis.[citation needed] CXCL9 predominantly mediates lymphocytic infiltration to the focal sites and suppresses tumor growth.[6]

For immune cell differentiation, some reports show that CXCL9 lead to Th1 polarization through CXCR3.[15] In vivo model by Zohar et al. showed that CXCL9, drove increased transcription of T-bet and RORγ, leading to the polarization of Foxp3− type 1 regulatory (Tr1) cells or T helper 17 (Th17) from naive T cells via STAT1, STAT4, and STAT5 phosphorylation.[15]

Several studies have shown that tumor associated macrophages (TAMs) play modulatory activities in the TME, and the CXCL9/CXCR3 axis impacts TAMs polarization. The TAMs have opposite effects; M1 for anti-tumor activities, and M2 for pro-tumor activities. Oghumu et al clarified that CXCR3 deficient mice displayed increased IL-4 production and M2 polarization in a murine breast cancer model, and decreased innate and immune cell-mediated anti-tumor responses.[16]

For immune cell activation, CXCL9 stimulate immune cells through Th1 polarization and activation. Th1 cells produce IFN-γ, TNF-α, IL-2 and enhance anti-tumor immunity by stimulating CTLs, NK cells and macrophages.[17] The IFN-γ-dependent immune activation loop also promotes CXCL9 release.[3]

IDO1: Indoleamine-pyrrole 2,3-dioxygenase (IDO or INDO EC 1.13.11.52) is a heme-containing enzyme physiologically expressed in a number of tissues and cells, such as the small intestine, lungs IDO is an important molecule in the mechanisms of tolerance and its physiological functions include the suppression of potentially dangerous inflammatory processes in the body.[15] IDO also plays a role in natural defense against microorganisms. Expression of IDO is induced by interferon-gamma, which explains why the expression increases during inflammatory diseases

IL.13: nterleukin 13 (IL-13) is a protein that in humans is encoded by the IL13 gene.[4][5][6] IL-13 was first cloned in 1993 and is located on chromosome 5q31 with a length of 1.4kb.[4] It has a mass of 13 kDa and folds into 4 alpha helical bundles.[7] The secondary structural features of IL-13 are similar to that of Interleukin 4 (IL-4); however it only has 25% sequence identity to IL-4 and is capable of IL-4 independent signaling.[7][4][8] IL-13 is a cytokine secreted by T helper type 2 (Th2) cells, CD4 cells, natural killer T cell, mast cells,

basophils, eosinophils and nuocytes.[7] Interleukin-13 is a central regulator in IgE synthesis, goblet cell hyperplasia, mucus hypersecretion, airway hyperresponsiveness, fibrosis and chitinase up-regulation.[7] It is a mediator of allergic inflammation and different diseases including asthma.[7] IL-13 specifically induces physiological changes in parasitized organs that are required to expel the offending organisms or their products. For example, expulsion from the gut of a variety of mouse helminths requires IL-13 secreted by Th2 cells. IL-13 induces several changes in the gut that create an environment hostile to the parasite, including enhanced contractions and glycoprotein hyper-secretion from gut epithelial cells, that ultimately lead to detachment of the organism from the gut wall and their removal.[13]

TNF: Tumor necrosis factor (TNF, cachexin, or cachectin; often called tumor necrosis factor alpha or TNF-α) is an adipokine and a cytokine. TNF is a member of the TNF superfamily, which consists of various transmembrane proteins with a homologous TNF domain. The primary role of TNF is in the regulation of immune cells. TNF, as an endogenous pyrogen, is able to induce fever, apoptotic cell death, cachexia, and inflammation, inhibit tumorigenesis and viral replication, and respond to sepsis via IL-1 and IL-6-producing cells. Dysregulation of TNF production has been implicated in a variety of human diseases including Alzheimer's disease,[12] cancer,[13] major depression,[14] psoriasis[15] and inflammatory bowel disease (IBD).[16] Though controversial, some studies have linked depression and IBD to increased levels of TNF.[17] [18] it is produced also by a broad variety of cell types including lymphoid cells, mast cells, endothelial cells, cardiac myocytes, adipose tissue, fibroblasts, and neurons.[51][unreliable medical source?] Large amounts of TNF are released in response to lipopolysaccharide, other bacterial products, and interleukin-1 (IL-1). In the skin, mast cells appear to be the predominant source of pre-formed TNF, which can be released upon inflammatory stimulus (e.g., LPS).[52]

TNF promotes the inflammatory response, which, in turn, causes many of the clinical problems associated with autoimmune disorders such as rheumatoid arthritis, ankylosing spondylitis, inflammatory bowel disease, psoriasis, hidradenitis suppurativa and refractory asthma. These disorders are sometimes treated by using a TNF inhibitor.

Group 3: IL1 RN: interleukin 1 receptor antagonist The interleukin-1 receptor antagonist (IL-1RA) is a protein that in humans is encoded by the IL1RN gene.[IL-1RA is a member of the interleukin 1 cytokine family. IL1Ra is secreted by various types of cells including immune cells, epithelial cells, and adipocytes, and is a natural inhibitor of the pro-inflammatory effect of IL1β.[8] This protein inhibits the activities of interleukin 1, alpha (IL1A) and interleukin 1, beta (IL1B), and modulates a variety of interleukin 1 related immune and inflammatory responses.

MPO: Myeloperoxidase (MPO) is a peroxidase enzyme that in humans is encoded by the MPO gene on chromosome 17.[5] MPO is most abundantly expressed in neutrophil granulocytes (a subtype of white blood cells), and produces hypohalous acids to carry out their antimicrobial activity, including hypochlorous acid, the sodium salt of which is the chemical in bleach.[5][6] It is a lysosomal protein stored in azurophilic granules of the neutrophil and released into the extracellular space during degranulation.[7] Neutrophil myeloperoxidase has a heme pigment, which causes its green color in secretions rich in neutrophils, such as mucus and sputum.[8] The green color contributed to its outdated name verdoperoxidase.

MPO is a member of the XPO subfamily of peroxidases and produces hypochlorous acid (HOCl) from hydrogen peroxide (H2O2) and chloride anion (Cl^-) (or hypobromous acid if Br- is present) during the neutrophil's respiratory burst. It requires heme as a cofactor. Furthermore, it oxidizes tyrosine to tyrosyl radical using hydrogen peroxide as an oxidizing agent.[10][14] Hypochlorous acid and tyrosyl radical are cytotoxic, so they are used by the neutrophil to kill bacteria and other pathogens.[15] However, this hypochlorous acid may also cause oxidative damage in host tissue. Moreover, MPO oxidation of apoA-I reduces HDL-mediated inhibition of apoptosis and inflammation.[16] In addition, MPO mediates protein nitrosylation and the formation of 3-chlorotyrosine and dityrosine crosslinks.[10]

Group 4: CASP1: Caspase-1/Interleukin-1 converting enzyme (ICE) is an evolutionarily conserved enzyme that proteolytically cleaves other proteins, such as the precursors of the inflammatory cytokines interleukin 1β and interleukin 18 as well as the pyroptosis inducer Gasdermin D, into active mature peptides.[5][6][7] It plays a

central role in cell immunity as an inflammatory response initiator. Once activated through formation of an inflammasome complex, it initiates a proinflammatory response through the cleavage and thus activation of the two inflammatory cytokines, interleukin 1β (IL-1β) and interleukin 18 (IL-18) as well as pyroptosis, a programmed lytic cell death pathway, through cleavage of Gasdermin D.[8] The two inflammatory cytokines activated by Caspase-1 are excreted from the cell to further induce the inflammatory response in neighboring cells.[9]

MUC5AC: Mucin 5AC (Muc5AC) is a protein that in humans is encoded by the MUC5AC gene.[5][6][7]

Muc5AC is a large gel-forming glycoprotein. In the respiratory tract it protects against infection by binding to inhaled pathogens that are subsequently removed by mucociliary clearance. Overproduction of Muc5AC can contribute to diseases such as asthma and chronic obstructive pulmonary disease,[8] and has also been associated with greater protection against influenza infection.[9]

Group 5: PRF1: Perforin-1 is a protein that in humans is encoded by the PRF1 gene and the Prf1 gene in mice. Perforin is a pore forming cytolytic protein found in the granules of cytotoxic T lymphocytes (CTLs) and natural killer cells (NK cells). Upon degranulation, perforin molecules translocate to the target cell with the help of calreticulin, which works as a chaperone protein to prevent perforin from degrading. Perforin then binds to the target cell's plasma membrane via membrane phospholipids while phosphatidylcholine binds calcium ions to increase perforin's affinity to the membrane.[8] Perforin oligomerises in a Ca2+ dependent manner to form pores on the target cell. The pore formed allows for the passive diffusion of a family of pro-apoptotic proteases, known as the granzymes, into the target cell.[9] The lytic membrane-inserting part of perforin is the MACPF domain.[10] This region shares homology with cholesterol-dependent cytolysins from Gram-positive bacteria.[11]

Perforin has structural and functional similarities to complement component 9 (C9). Like C9, this protein creates transmembrane tubules and is capable of lysing non-specifically a variety of target cells. This protein is one of the main cytolytic proteins of cytolytic granules, and it is known to be a key effector molecule for T-cell- and natural killer-cell-mediated cytolysis.[7] Perforin is thought to act by creating holes in the plasma membrane which triggers an influx of calcium and initiates membrane repair mechanisms. These repair mechanisms bring perforin and granzymes into early endosomes.[12]

CXCR3: Chemokine receptor CXCR3 is a Gαi protein-coupled receptor in the CXC chemokine receptor family. Other names for CXCR3 are G protein-coupled receptor 9 (GPR9) and CD183. There are three isoforms of CXCR3 in humans: CXCR3-A, CXCR3-B and chemokine receptor 3-alternative (CXCR3-alt).[5] CXCR3-A binds to the CXC chemokines CXCL9 (MIG), CXCL10 (IP-10), and CXCL11 (I-TAC)[6] whereas CXCR3-B can also bind to CXCL4 in addition to CXCL9, CXCL10, and CXCL11.[7] CXCR3 is expressed primarily on activated T lymphocytes and NK cells,[8] and some epithelial cells. CXCR3 and CCR5 are preferentially expressed on Th1 cells, whereas Th2 cells favor the expression of CCR3 and CCR4. CXCR3 ligands that attract Th1 cells can concomitantly block the migration of Th2 cells in response to CCR3 ligands, thus enhancing the polarization of effector T cell recruitment.

MYD88: Myeloid differentiation primary response 88 (MYD88) is a protein that, in humans, is encoded by the MYD88 gene. Model organisms have been used in the study of MYD88 function. The gene was originally discovered and cloned by Dan Liebermann and Barbara Hoffman in mice.[7] In that species it is a universal adapter protein as it is used by almost all TLRs (except TLR 3) to activate the transcription factor NF-κB. Mal (also known as TIRAP) is necessary to recruit Myd88 to TLR 2 and TLR 4, and MyD88 then signals through IRAK.[8] It also interacts functionally with amyloid formation and behavior in a transgenic mouse model of Alzheimer's disease.[9]

Myd88 knockout mouse phenotype A conditional knockout mouse line, called Myd88tm1a(EUCOMM)Wtsi[13][14] was generated as part of the International Knockout Mouse Consortium program — a high-throughput mutagenesis project to generate and distribute animal models of disease to interested scientists.[15][16][17] Male and female animals underwent a standardized phenotypic screen to determine the effects of deletion.[11][18] Twenty-one tests were carried out on homozygous mutant animals, revealing one abnormality: male

mutants had an increased susceptibility to bacterial infection. The MYD88 gene provides instructions for making a protein involved in signaling within immune cells. The MyD88 protein acts as an adapter, connecting proteins that receive signals from outside the cell to the proteins that relay signals inside the cell. In innate immunity, the MyD88 plays a pivotal role in immune cell activation through Toll-like receptors (TLRs), which belong to large group of pattern recognition receptors (PRR). In general, these receptors sense common patterns which are shared by various pathogens – Pathogen-associated molecular pattern (PAMPs), or which are produced/released during cellular damage – damage-associated molecular patterns (DAMPs).[19]

IL.6: Interleukin 6 (IL-6) is an interleukin that acts as both a pro-inflammatory cytokine and an anti-inflammatory myokine. In humans, it is encoded by the IL6 gene.[5]

In addition, osteoblasts secrete IL-6 to stimulate osteoclast formation. Smooth muscle cells in the tunica media of many blood vessels also produce IL-6 as a pro-inflammatory cytokine. IL-6's role as an anti-inflammatory myokine is mediated through its inhibitory effects on TNF-alpha and IL-1 and its activation of IL-1ra and IL-10.

Immune system IL-6 is secreted by macrophages in response to specific microbial molecules, referred to as pathogen-associated molecular patterns (PAMPs). These PAMPs bind to an important group of detection molecules of the innate immune system, called pattern recognition receptors (PRRs), including Toll-like receptors (TLRs). These are present on the cell surface and intracellular compartments and induce intracellular signaling cascades that give rise to inflammatory cytokine production. IL-6 is an important mediator of fever and of the acute phase response.

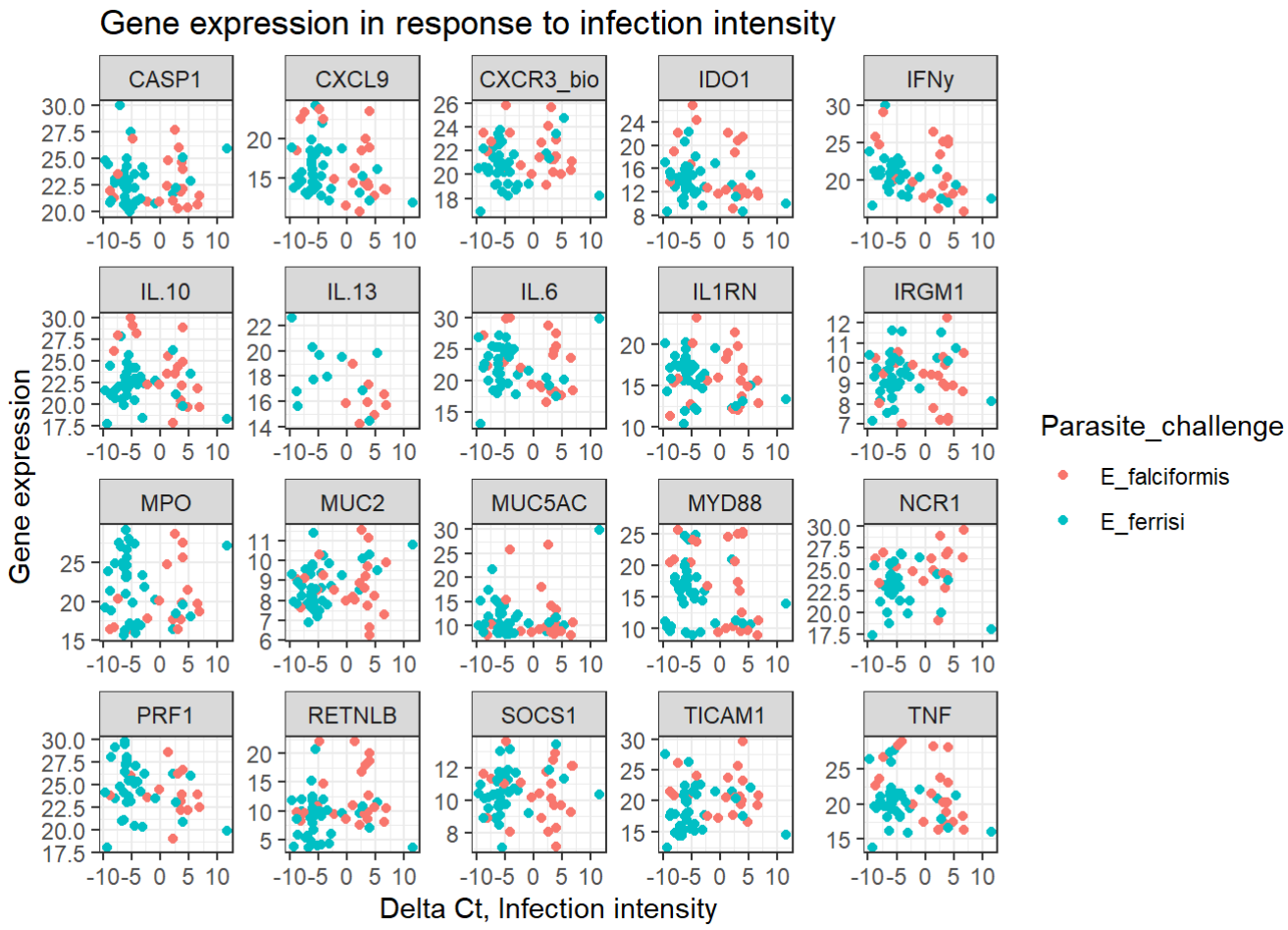
IL-6 is responsible for stimulating acute phase protein synthesis, as well as the production of neutrophils in the bone marrow. It supports the growth of B cells and is antagonistic to regulatory T cells.

NCR1: Natural cytotoxicity triggering receptor 1 is a protein that in humans is encoded by the NCR1 gene.[

RETNL:

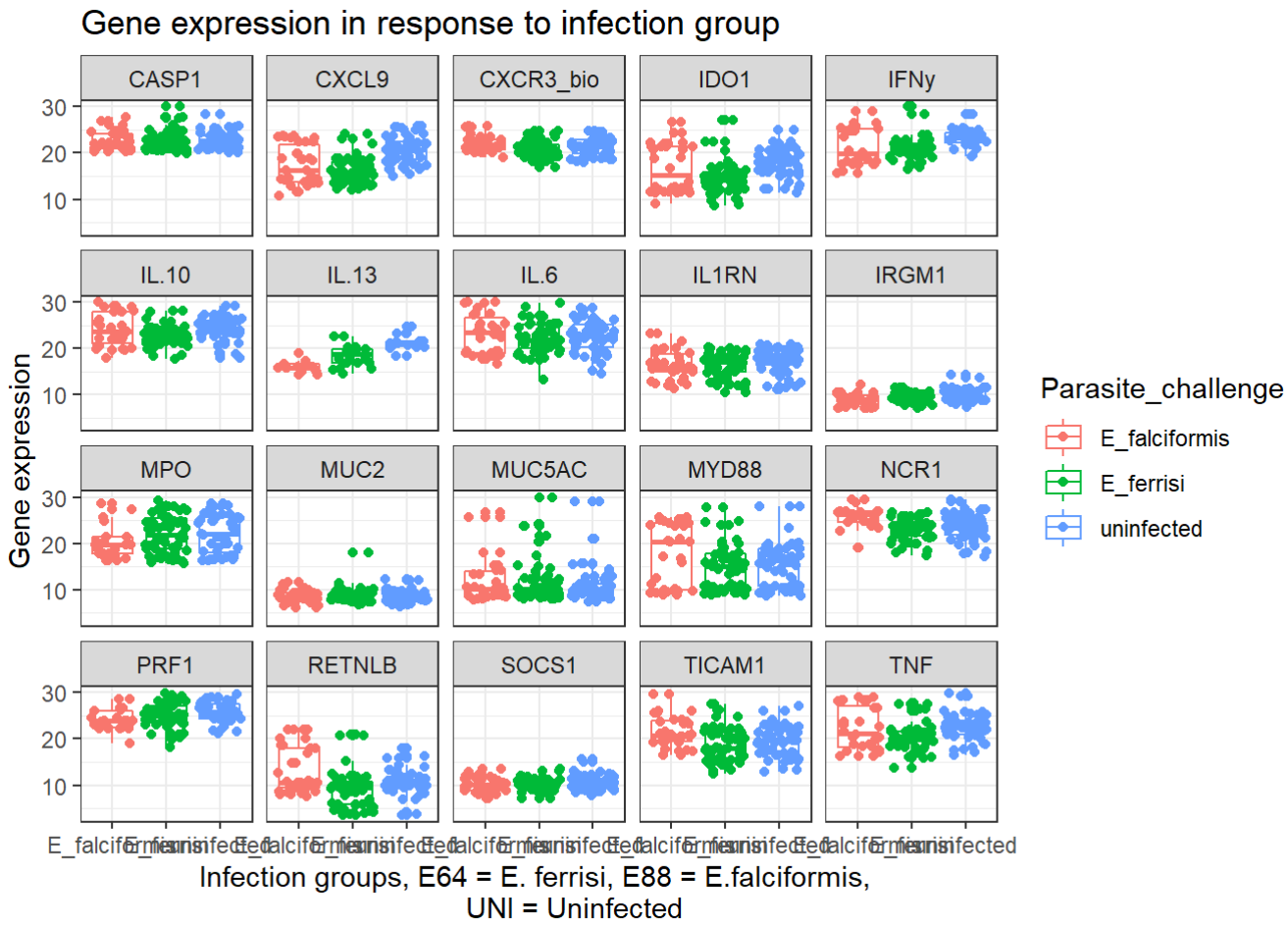
TICAM1: TIRP is a Toll/interleukin-1 receptor (IL1R; MIM 147810) (TIR) domain-containing adaptor protein involved in Toll receptor signaling

```
## Warning: Removed 121 rows containing missing values (geom_point).
```



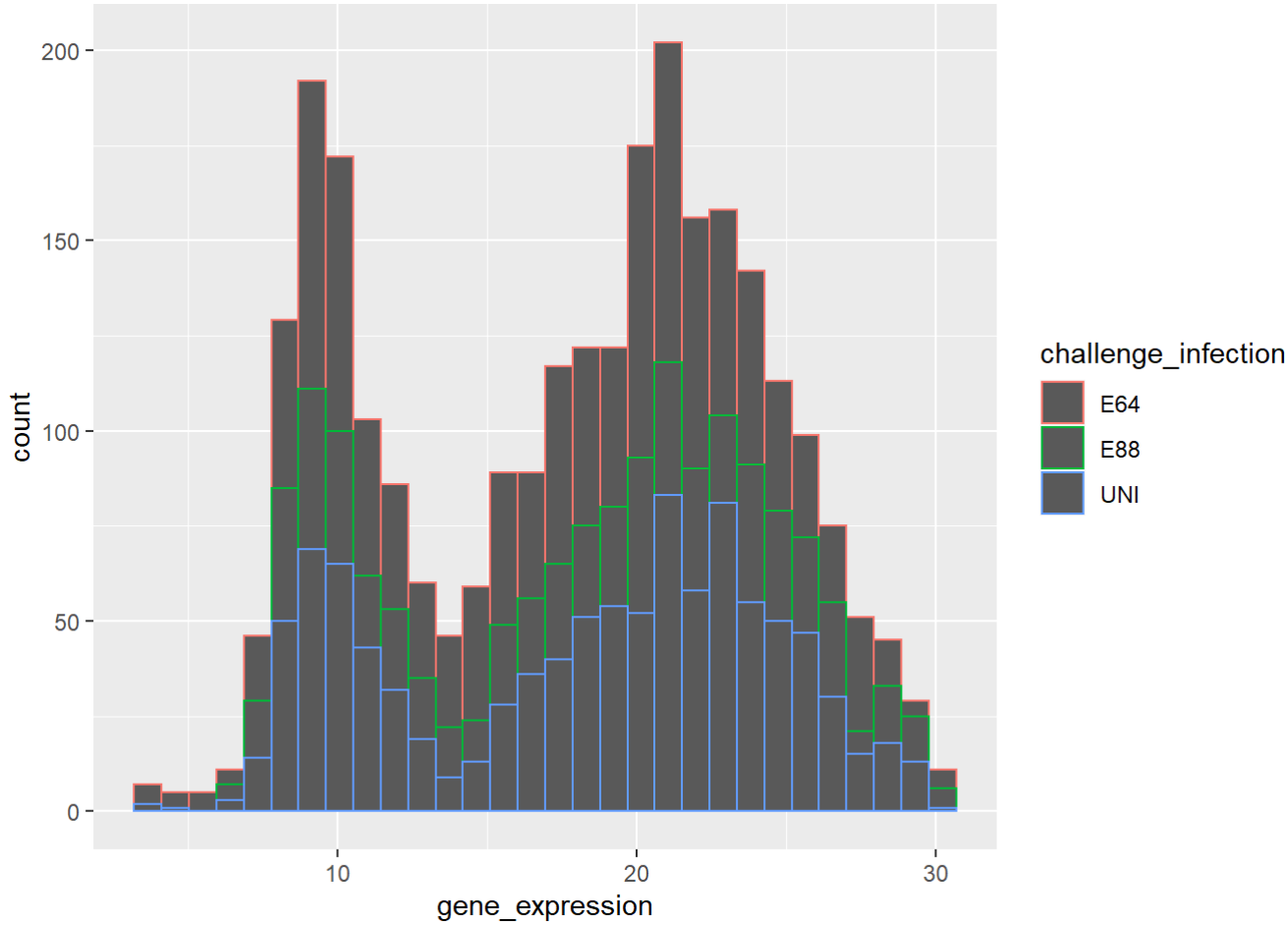

```
## Warning: Removed 244 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 244 rows containing missing values (geom_point).
```




```
## Warning: Ignoring unknown parameters: echo
```

```
## Warning: Removed 244 rows containing non-finite values (stat_bin).
```



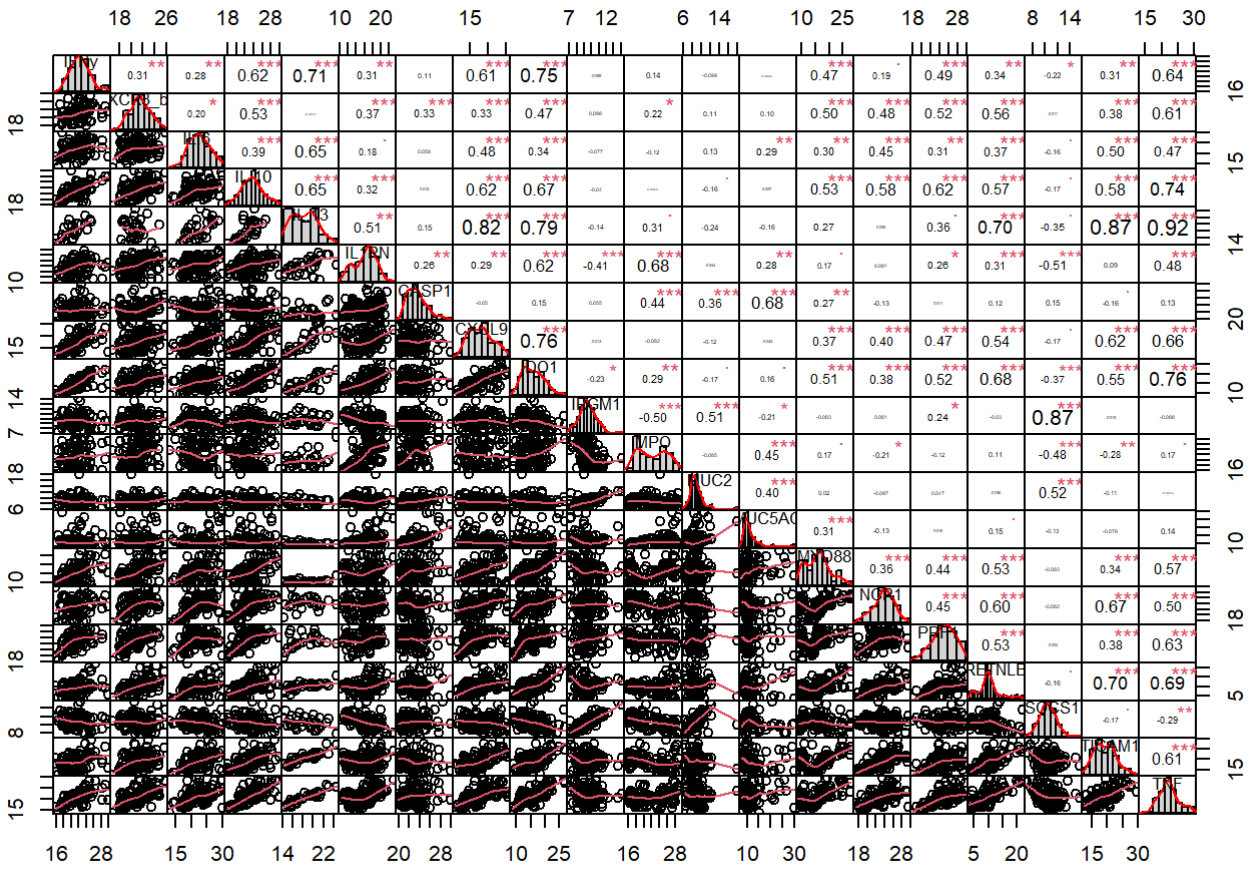
It is possible to compute a pca with missing data using the package missMDA. The missMDA package is dedicated to missing values in exploratory multivariate data analysis: single imputation/multiple imputation, etc.

Following the tutorial of the package author: Francois Husson: https://www.youtube.com/watch?v=OOM8_FH6_8o (https://www.youtube.com/watch?v=OOM8_FH6_8o)

3. PCA

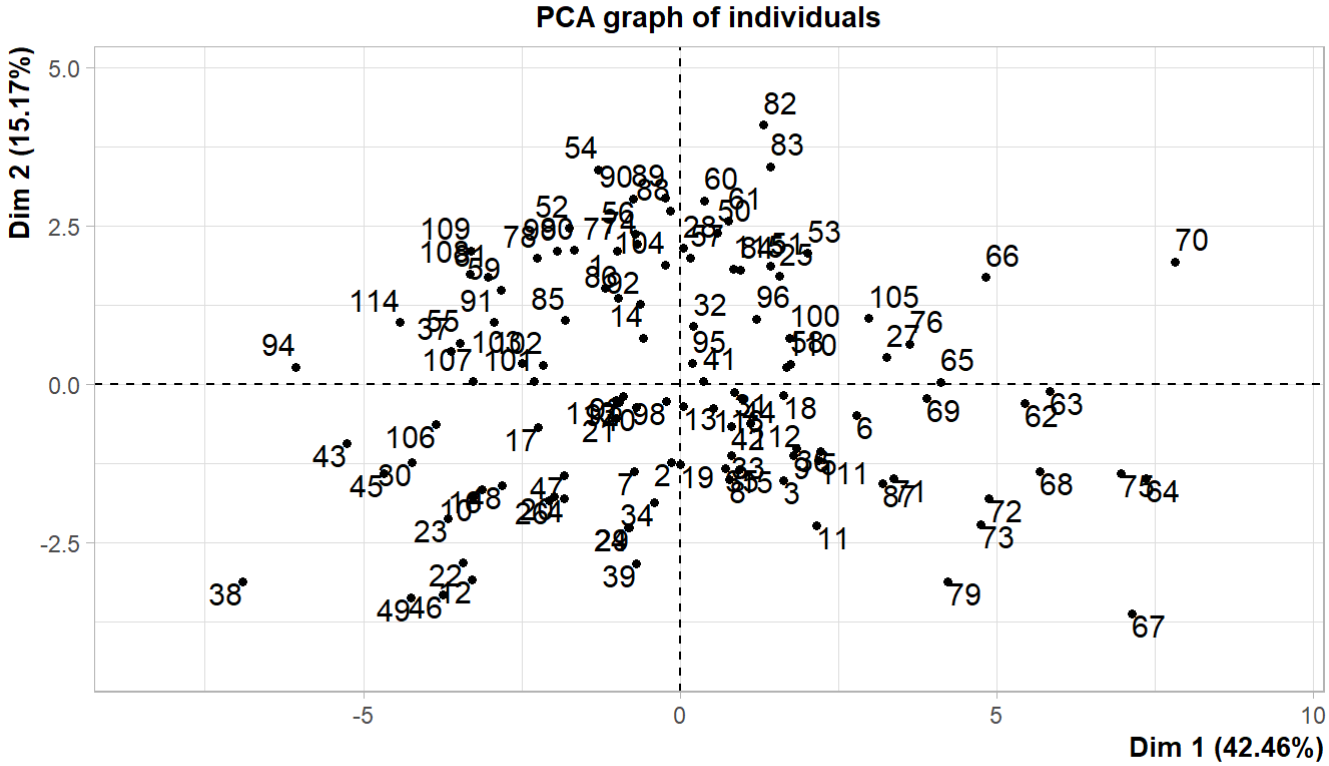
Handling missing data in a pca:

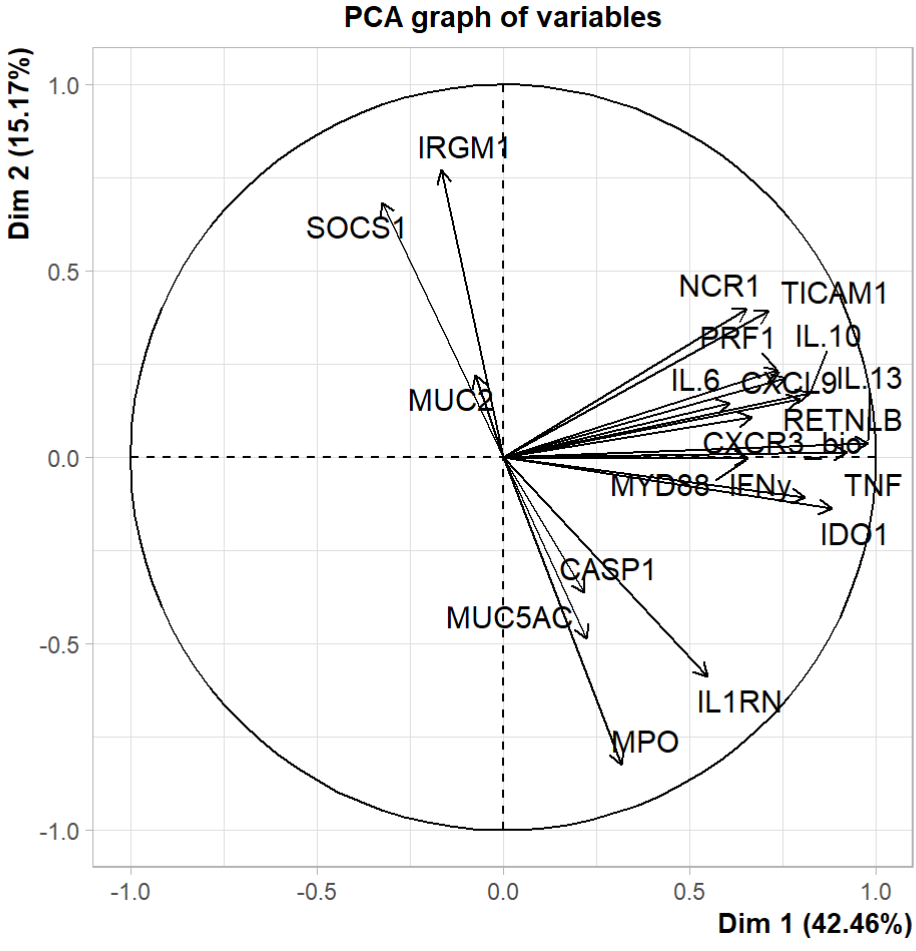
Bad methods: removing individuals with missing data or replacing missing data with the mean (default setting in many packages).



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 individuals or variables.





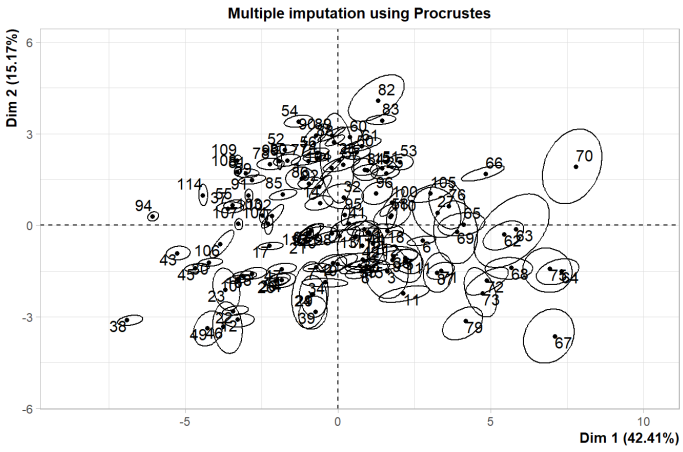
Caution: When imputing data, the percentages of inertia associated with the first dimensions will be overestimated.

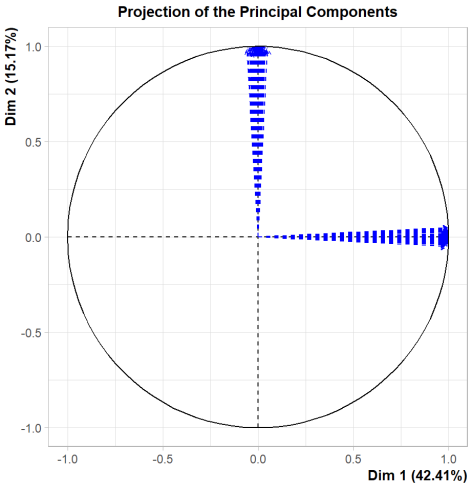
Another problem: the imputed data are, when the pca is performed considered like real observations. But they are estimations!!

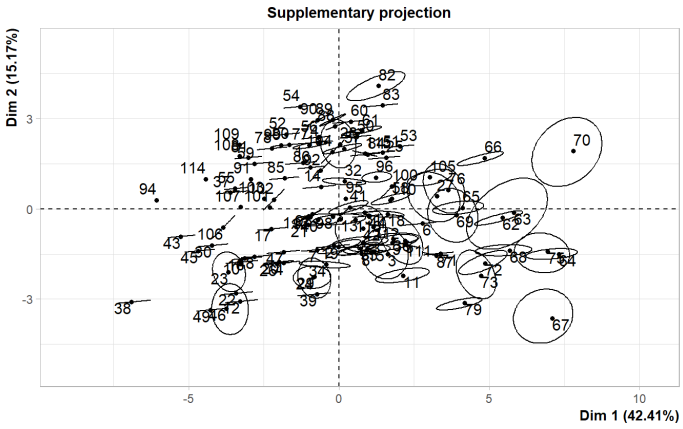
Visualizing uncertainty due to issing data:

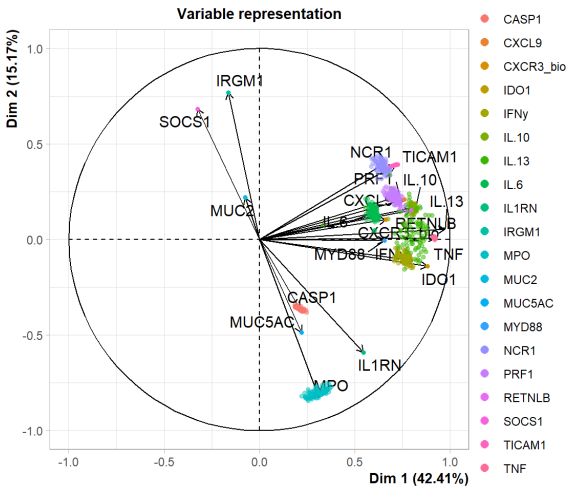
–> mulrimple imputation: generate several plausible values for each missing data point

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

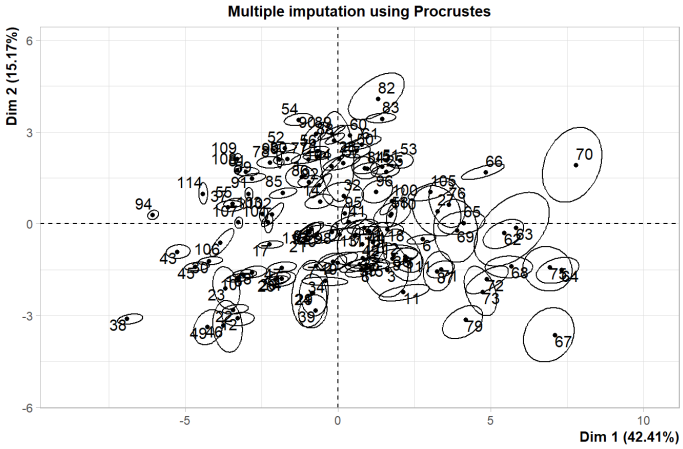




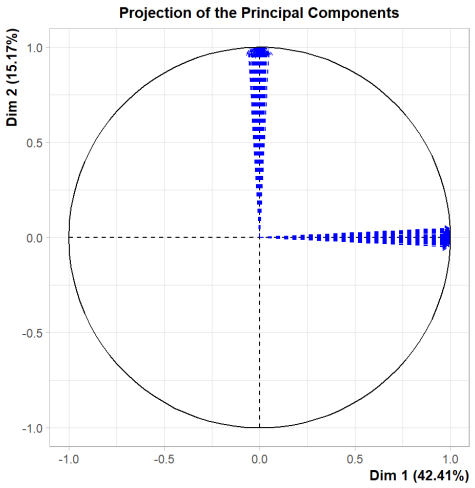




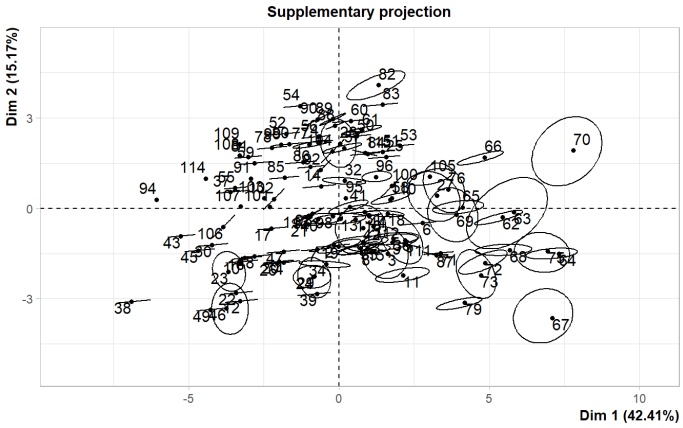

```
## $PlotIndProc
```



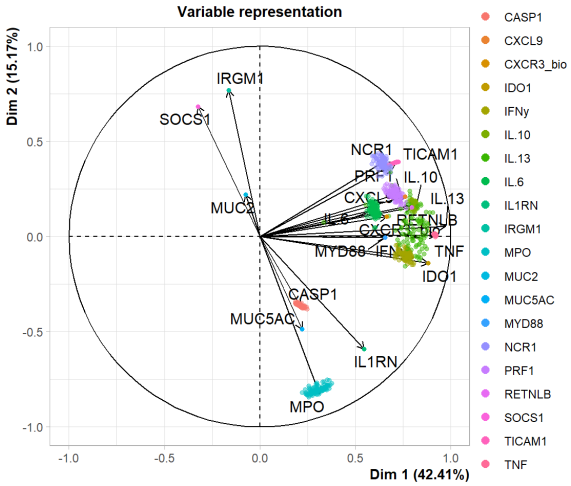

```
##
## $PlotDim
```




```
##
## $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

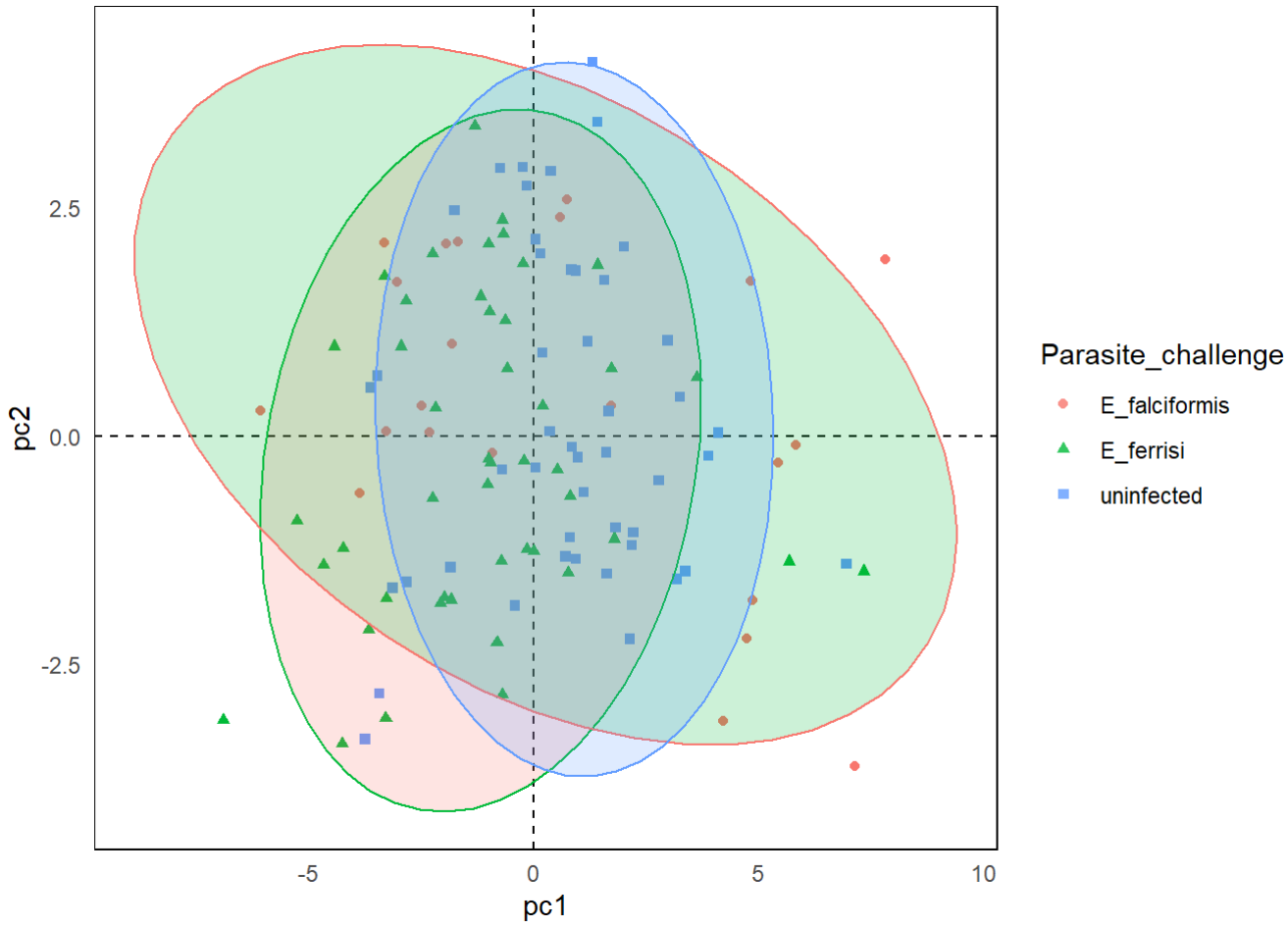
Therefore in the last plots every individual is getting an eclipse as they are as well influenced by the missing data of the others.

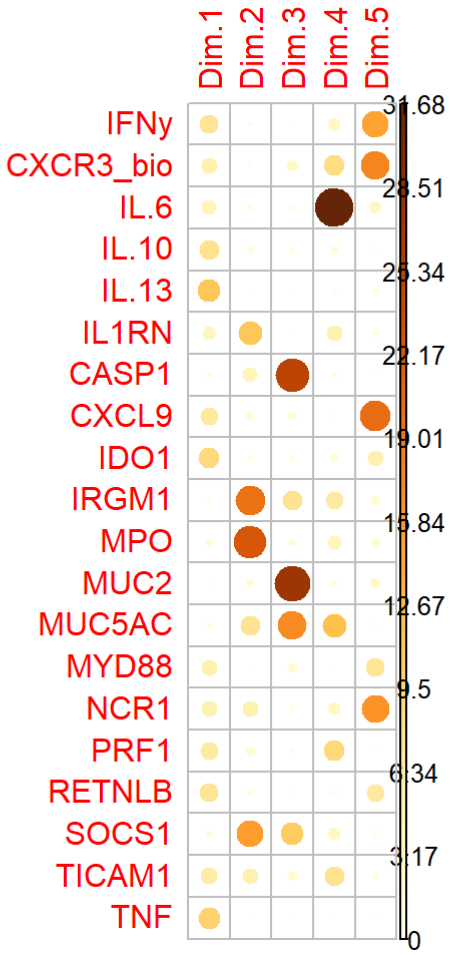
THe plot with the dimensions shows the projections of the pca dimensions of each imputed table on the pca plane obtained using the original imputed data table

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.

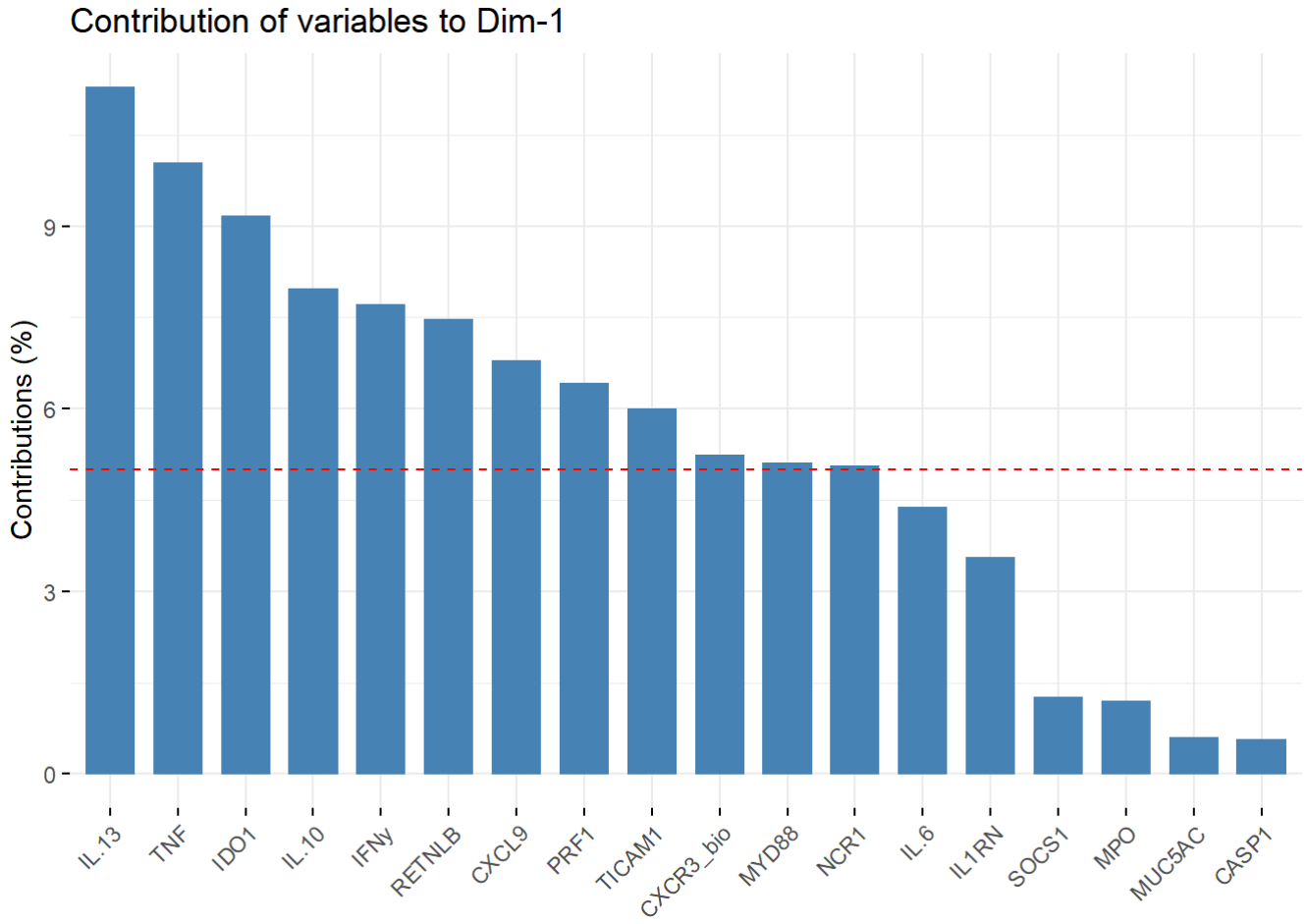
Biplot of the imputed gene pca

```
#Now we can make our initial plot of the PCA.
imputed_expr %>%
  pivot_longer(cols = all_of(Genes), names_to = "Gene", values_to = "gene_expression")  %>%
  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 = FA
LSE,
               level = 0.95) +
  theme_minimal() +
  theme(panel.grid = element_blank(), panel.border = element_rect(fill= "transparent"))
```

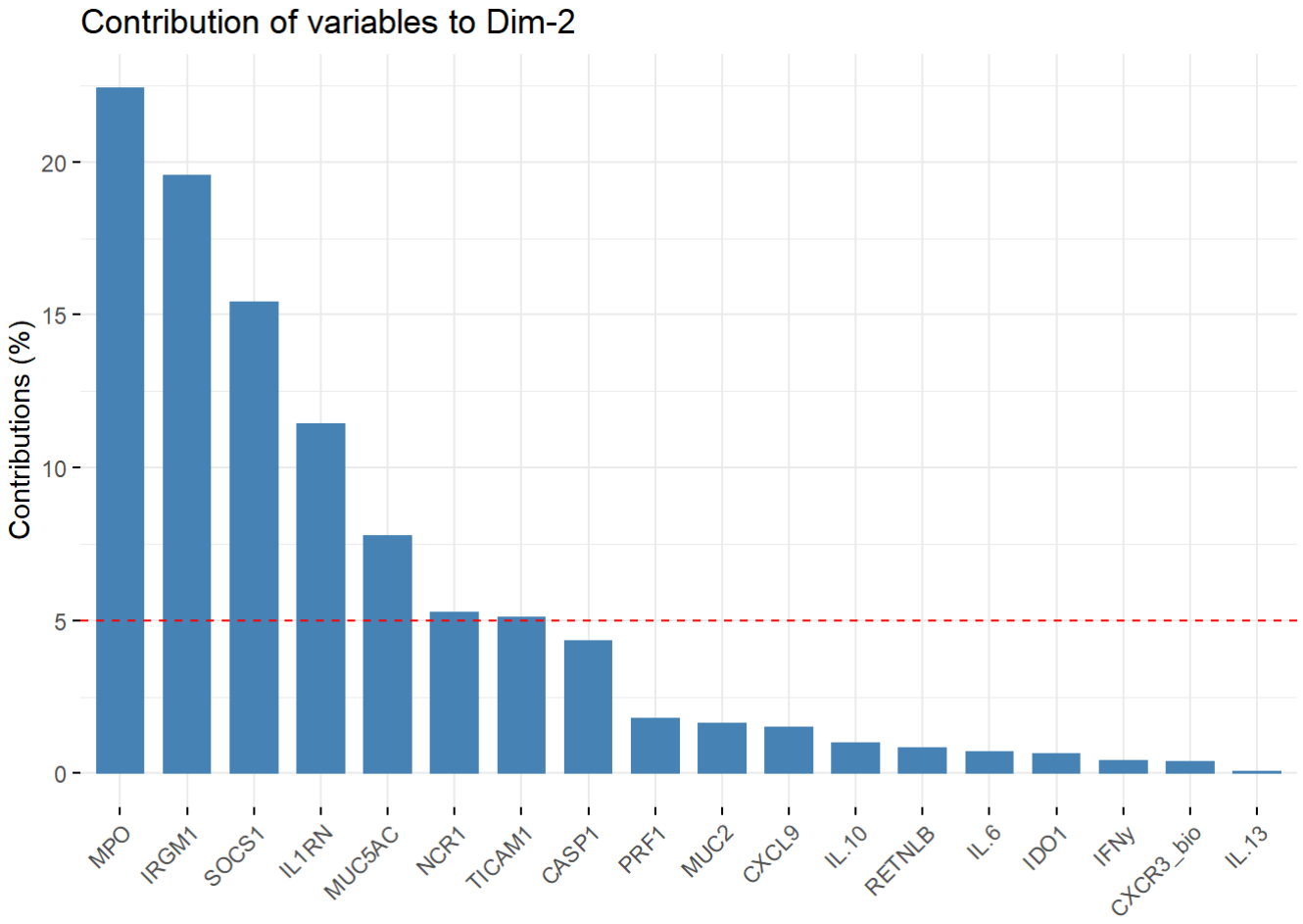


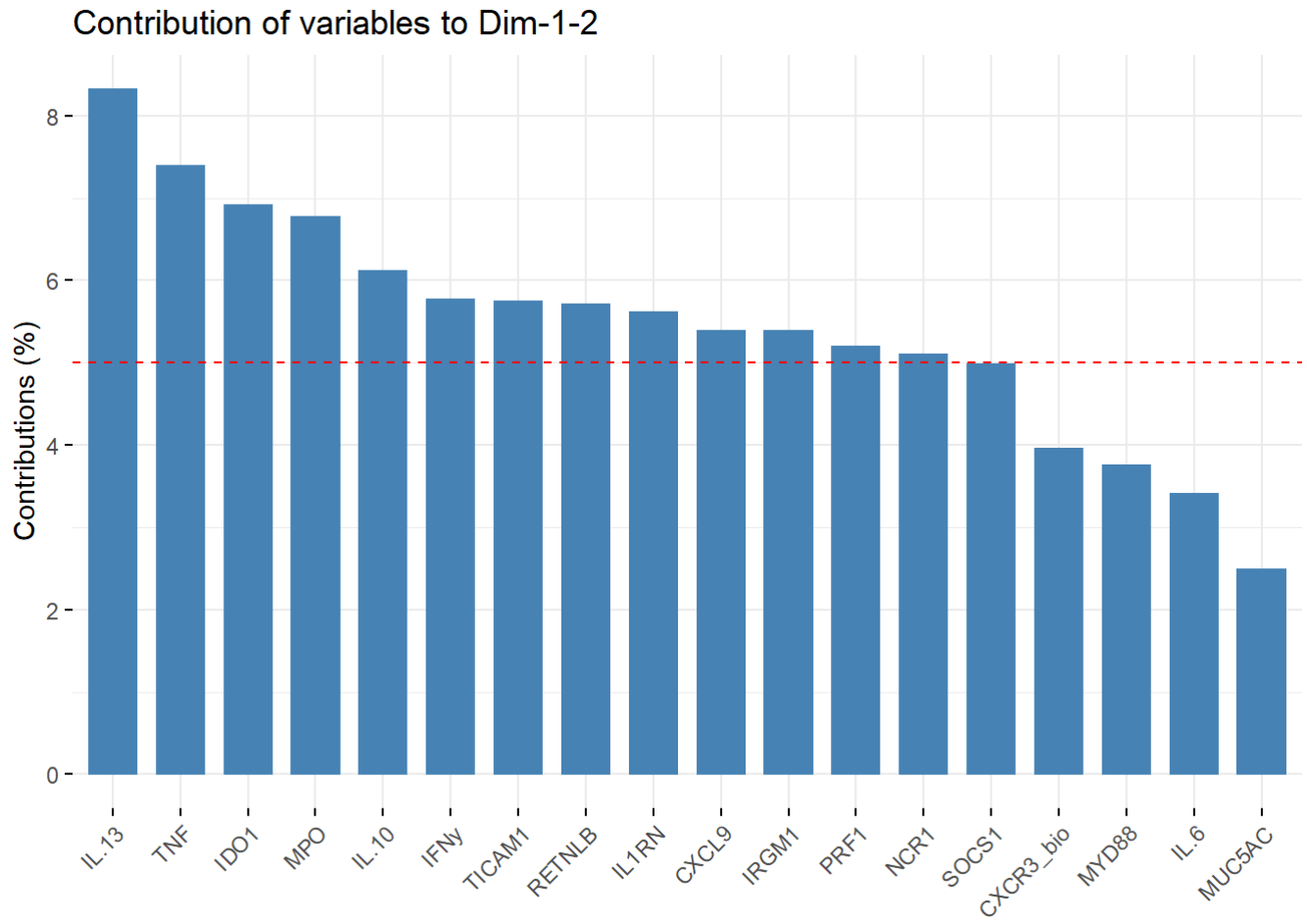


The function fviz_contrib() [factoextra package] can be used to draw a bar plot of variable contributions. If your data contains many variables, you can decide to show only the top contributing variables. The R code below shows the top 10 variables contributing to the principal components:



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

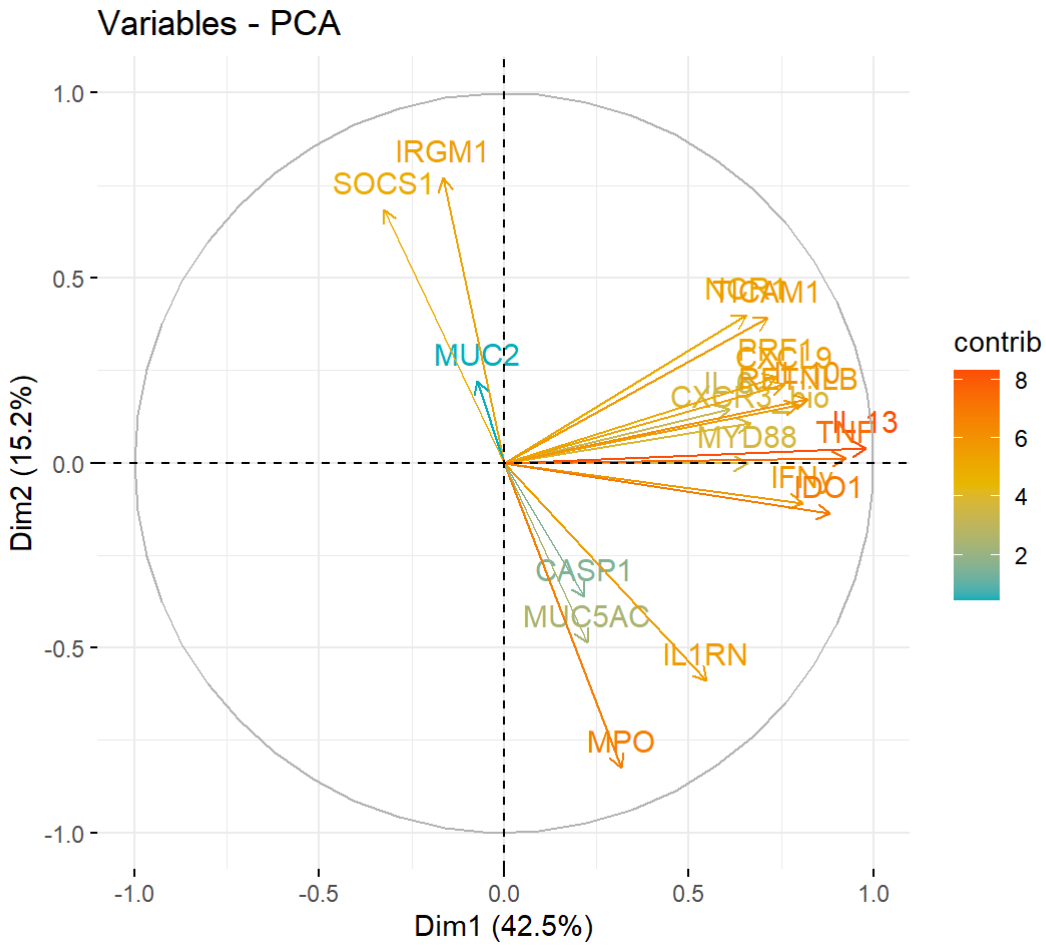




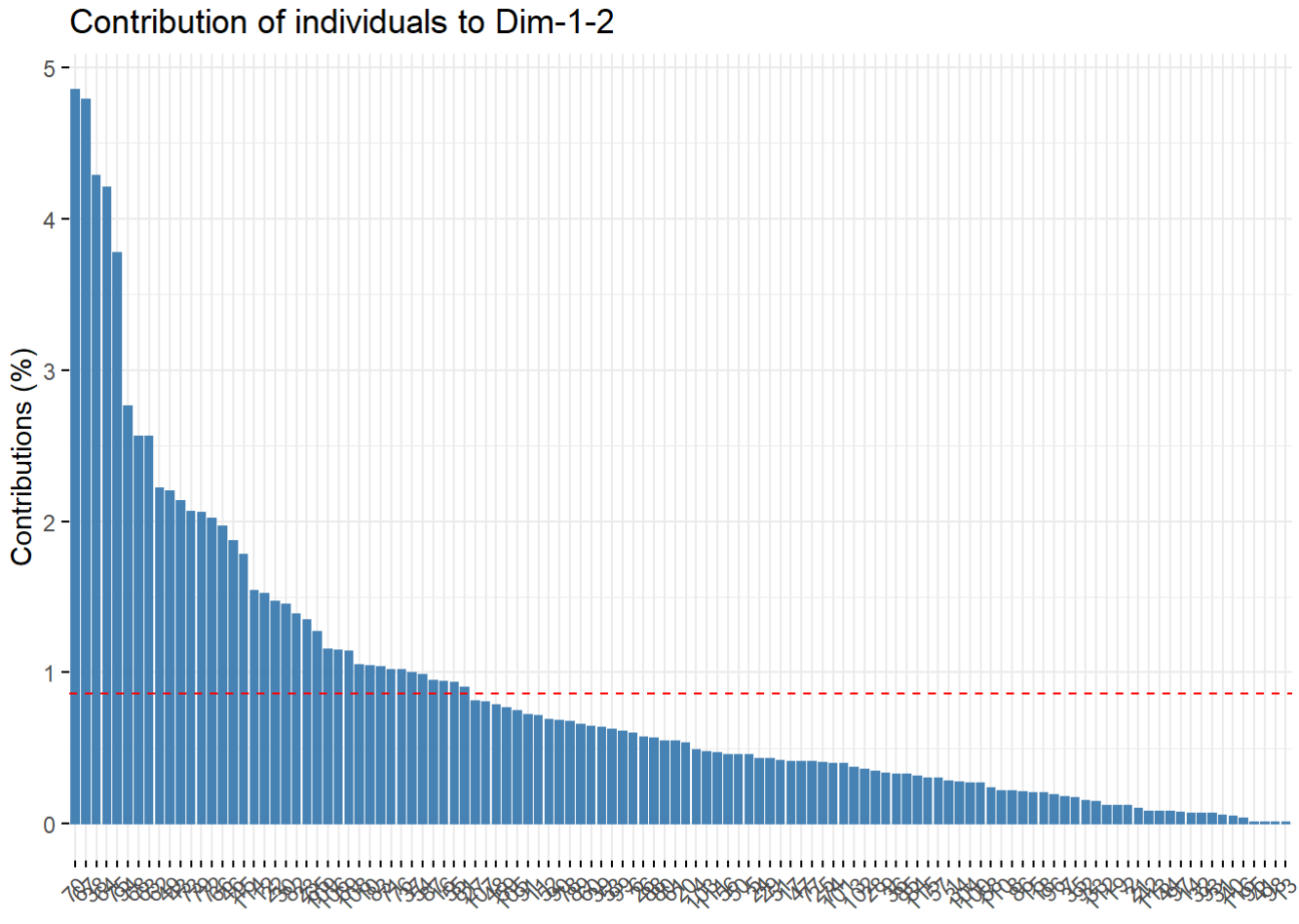
The red dashed line on the graph above indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would be 1/length(variables) = 1/10 = 10%. For a given component, a variable with a contribution larger than this cutoff could be considered as important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

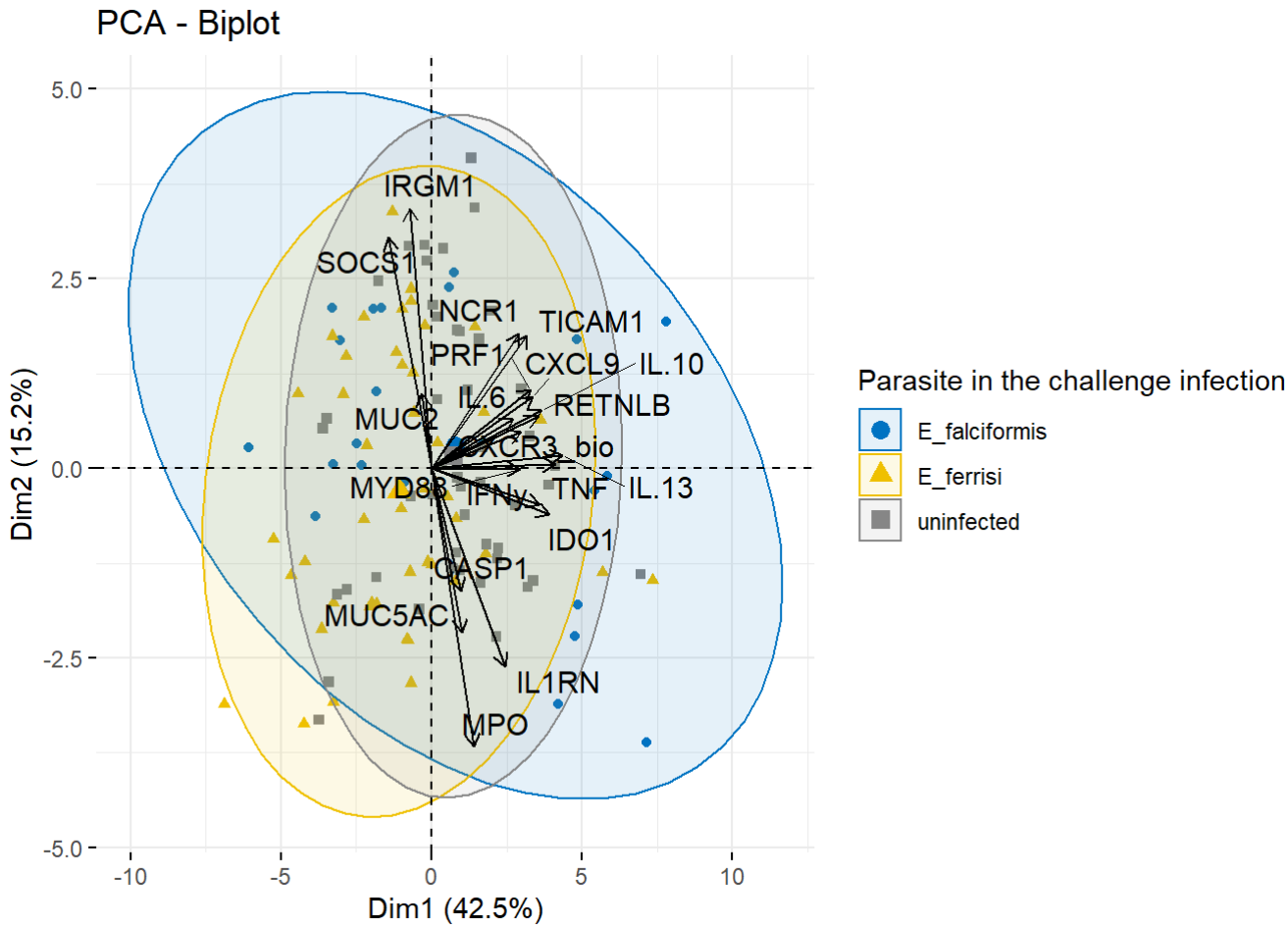
C1 and C2 are the contributions of the variable on PC1 and PC2, respectively Eig1 and Eig2 are the eigenvalues of PC1 and PC2, respectively. Recall that eigenvalues measure the amount of variation retained by each PC. In this case, the expected average contribution (cutoff) is calculated as follow: As 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 = max_WL ~ pc1 + pc2 + Parasite_challenge, data = imputed_expr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4014  -3.0944   0.1175   3.5262  14.3050
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   85.6719     1.1323  75.664  < 2e-16 ***
## pc1                            0.1272     0.1762   0.722   0.4718
## pc2                           -0.7278     0.2834  -2.568   0.0116 *
## Parasite_challengeE_ferrisi    6.0895     1.4092   4.321 3.40e-05 ***
## Parasite_challengeuninfected  10.4534     1.3576   7.700 6.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.253 on 111 degrees of freedom
## Multiple R-squared:  0.3818, Adjusted R-squared:  0.3596
## F-statistic: 17.14 on 4 and 111 DF,  p-value: 5.657e-11
```

```
## [1] 720.9133
```

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + Parasite_challenge + hybrid_status,
##     data = imputed_expr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7141  -3.5997   0.4672   3.5380  13.9501
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      86.0601     1.3838  62.193  < 2e-16 ***
## pc1                               0.1364     0.2199   0.620   0.5364
## pc2                              -0.5959     0.3144  -1.896   0.0607 .
## Parasite_challengeE_ferrisi       5.9059     1.4538   4.062 9.34e-05 ***
## Parasite_challengeuninfected     10.0684     1.4516   6.936 3.30e-10 ***
## hybrid_statusF0 M. m. musculus   -1.1985     1.4579  -0.822   0.4129
## hybrid_statusF1 hybrid            1.4620     1.6821   0.869   0.3867
## hybrid_statusF1 M. m. domesticus -1.7765     2.2126  -0.803   0.4238
## hybrid_statusF1 M. m. musculus    1.7684     2.6843   0.659   0.5115
## hybrid_statusother               -0.3591     1.4859  -0.242   0.8095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.288 on 106 degrees of freedom
## Multiple R-squared:  0.4017, Adjusted R-squared:  0.3509
## F-statistic: 7.907 on 9 and 106 DF,  p-value: 7.337e-09
```

```
## [1] 727.1249
```

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

https://www.rdocumentation.org/packages/lmtest/versions/0.9-38/topics/lrtest (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(max_WL ~ pc1 + pc2 + hybrid_status, data = imputed_expr)
weight_no_pc1 <- lm(max_WL ~ pc2 + hybrid_status, data = imputed_expr)
weight_no_pc2 <- lm(max_WL ~ pc1  + hybrid_status, data = imputed_expr)
weight_no_hybrid <- lm(max_WL ~ pc1 + pc2, data = imputed_expr)
lrtest(weight_lm3, weight_no_pc1)
```

```
## Likelihood ratio test
##
## Model 1: max_WL ~ pc1 + pc2 + hybrid_status
## Model 2: max_WL ~ pc2 + hybrid_status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -374.55
## 2   8 -375.87 -1 2.6379     0.1043
```

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

```
## Likelihood ratio test
##
## Model 1: max_WL ~ pc1 + pc2 + hybrid_status
## Model 2: max_WL ~ pc1 + hybrid_status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -374.55
## 2   8 -374.96 -1 0.8221     0.3646
```

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

```
## Likelihood ratio test
##
## Model 1: max_WL ~ pc1 + pc2 + hybrid_status
## Model 2: max_WL ~ pc1 + pc2
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -374.55
## 2   4 -379.64 -5 10.186    0.07014 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + hybrid_status, data = imputed_expr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.942  -3.138   0.991   4.739   9.868
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      92.5229     1.0319  89.664   <2e-16 ***
## pc1                               0.3934     0.2496   1.576   0.1179
## pc2                              -0.3243     0.3701  -0.876   0.3827
## hybrid_statusF0 M. m. musculus   -1.1490     1.7436  -0.659   0.5113
## hybrid_statusF1 hybrid            3.7568     1.9749   1.902   0.0598 .
## hybrid_statusF1 M. m. domesticus -0.3187     2.6314  -0.121   0.9038
## hybrid_statusF1 M. m. musculus    3.9912     3.1916   1.251   0.2138
## hybrid_statusother               -2.5944     1.7376  -1.493   0.1383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.332 on 108 degrees of freedom
## Multiple R-squared:  0.1259, Adjusted R-squared:  0.06928
## F-statistic: 2.223 on 7 and 108 DF,  p-value: 0.03774
```

```
## [1] 767.095
```

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2 + infection_history, data = imputed_expr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4909  -3.4963   0.2167   3.0235  14.3776
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           90.04699    1.95916  45.962  < 2e-16
## pc1                                    0.02434    0.17334   0.140  0.88862
## pc2                                   -0.61060    0.29415  -2.076  0.04035
## infection_historyfalciformis_ferrisi   2.11290    2.34416   0.901  0.36947
## infection_historyfalciformis_uninfected  6.75252    2.36949   2.850  0.00527
## infection_historyferrisi_falciformis  -7.57149    2.57368  -2.942  0.00401
## infection_historyferrisi_ferrisi       2.81356    2.33435   1.205  0.23080
## infection_historyferrisi_uninfected    5.45339    2.19195   2.488  0.01442
## infection_historyuninfected            7.17664    2.60969   2.750  0.00702
## infection_historyuninfected_falciformis -4.51846    2.83429  -1.594  0.11389
## infection_historyuninfected_ferrisi   -2.60534    2.67011  -0.976  0.33144
##
## (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: 5.021 on 105 degrees of freedom
## Multiple R-squared:  0.4656, Adjusted R-squared:  0.4147
## F-statistic: 9.149 on 10 and 105 DF,  p-value: 1.007e-10
```

```
## [1] 716.0163
```

```
##
## Call:
## lm(formula = max_WL ~ pc1 + pc2, data = imputed_expr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.913  -3.236   1.379   5.127  10.471
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  92.3746     0.6006 153.811   <2e-16 ***
## pc1           0.1702     0.2061   0.826   0.4107
## pc2          -0.7501     0.3448  -2.175   0.0317 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.468 on 113 degrees of freedom
## Multiple R-squared:  0.04572,    Adjusted R-squared:  0.02883
## F-statistic: 2.707 on 2 and 113 DF,  p-value: 0.07108
```

```
##                    df      AIC
## weight_lm           6 720.9133
## weight_lm_exp_only  4 767.2808
```

repeating the heatmap on the now imputed data

```
gene <- imputed_expr %>% dplyr::select(c(EH_ID, all_of(Genes)))

 # 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))

 #switch the matrix back to a data frame format
 gene <- as.data.frame(gene)

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


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

```
##
## FALSE
##    20
```

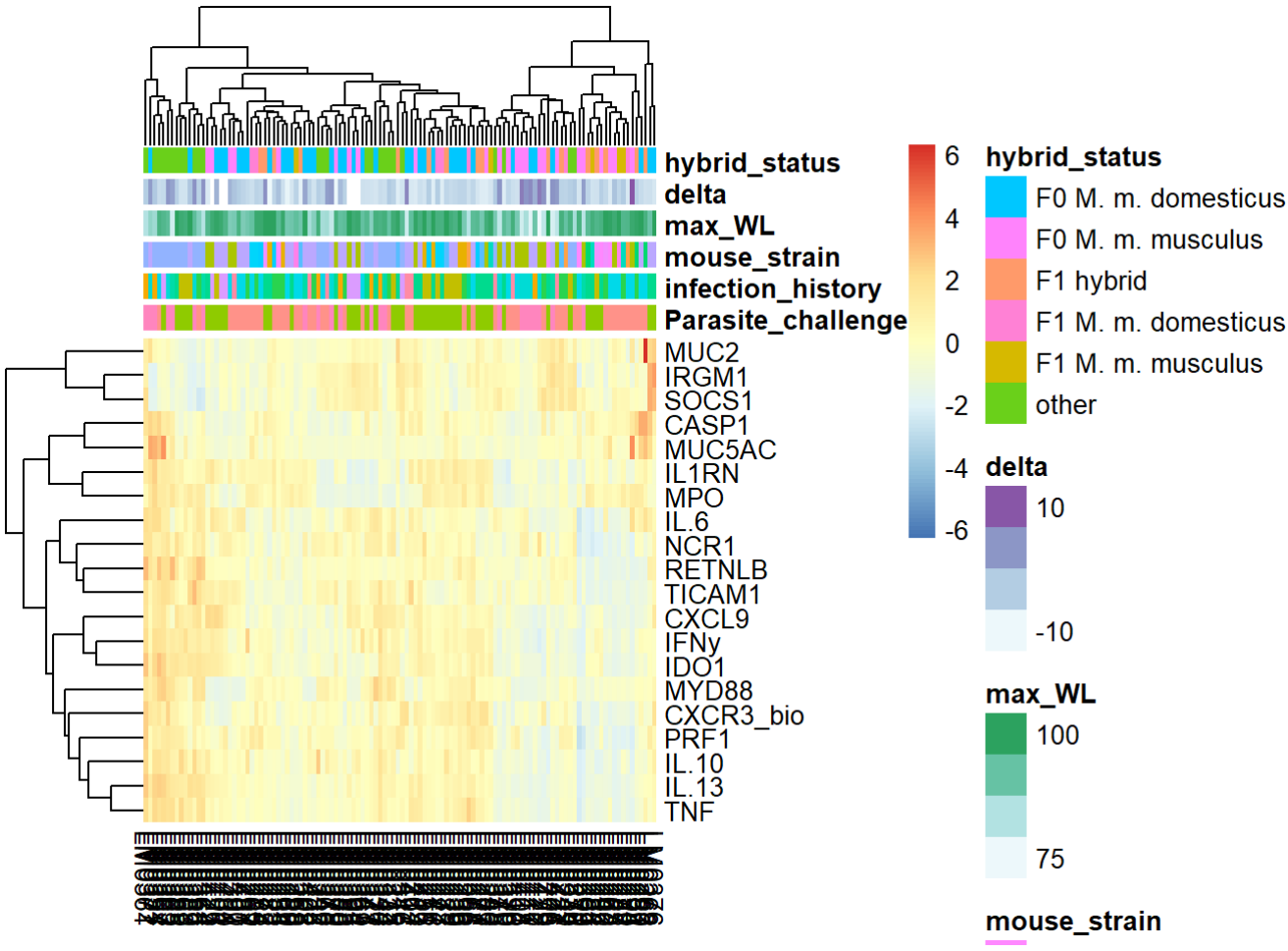
```
# turn the columns to numeric other wise the heatmap function will not work
heatmap_data[] <- lapply(heatmap_data, function(x) as.numeric(as.character(x)))

# 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)]

rownames(annotation_df) <- colnames(heatmap_data)
```

Heatmap on gene expression data:




```
write.csv(imputed_expr, "output_data/gene_expression/data_products/imputed_gene_expression.csv", row.names = FALSE)

write.csv(g2, "output_data/gene_expression/data_products/clean_gene_expression.csv", row.names = FALSE)
```