

# 4.Mice\_imputation\_field.rmd

Fay

2022-11-01

field

## Load libraries

```
library(mice)

##
## Attaching package: 'mice'
## The following object is masked from 'package:stats':
##
##   filter
## The following objects are masked from 'package:base':
##
##   cbind, rbind

library(tidyr)
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## v purrr   0.3.5
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks mice::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()

library(VIM)

## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
##
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
##
## The following object is masked from 'package:datasets':
##
##   sleep
```

```
library(fitdistrplus)

## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Loading required package: survival

library(fiturf)

##
## Attaching package: 'fiturf'
##
## The following object is masked from 'package:purrr':
##
##     rdunif

library(visdat)
```

## Load data

## Import data

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

# Vectors for selecting genes

#Lab genes
# The measurements of IL.12 and IRG6 are done with an other assay and will
#ignore for now
Gene_lab <- c("IFNy", "CXCR3", "IL.6", "IL.13", "IL.10",
              "IL1RN", "CASP1", "CXCL9", "IDO1", "IRGM1", "MPO",
              "MUC2", "MUC5AC", "MYD88", "NCR1", "PRF1", "RETNLB", "SOCS1",
              "TICAM1", "TNF") # "IL.12", "IRG6")

Genes_wild <- c("IFNy", "CXCR3", "IL.6", "IL.13", "IL.10",
               "IL1RN", "CASP1", "CXCL9", "IDO1", "IRGM1", "MPO",
               "MUC2", "MUC5AC", "MYD88", "NCR1", "PRF1", "RETNLB", "SOCS1",
               "TICAM1", "TNF") #, "IL.12", "IRG6")

Facs_lab <- c("CD4", "Treg", "Div_Treg", "Treg17", "Th1",
             "Div_Th1", "Th17", "Div_Th17", "CD8", "Act_CD8",
             "Div_Act_CD8", "IFNy_CD4", "IFNy_CD8", "Treg_prop",
             "IL17A_CD4")

Facs_wild <- c("Treg", "CD4", "Treg17", "Th1", "Th17", "CD8",
              "Act_CD8", "IFNy_CD4", "IL17A_CD4", "IFNy_CD8")
```

# Field data imputation

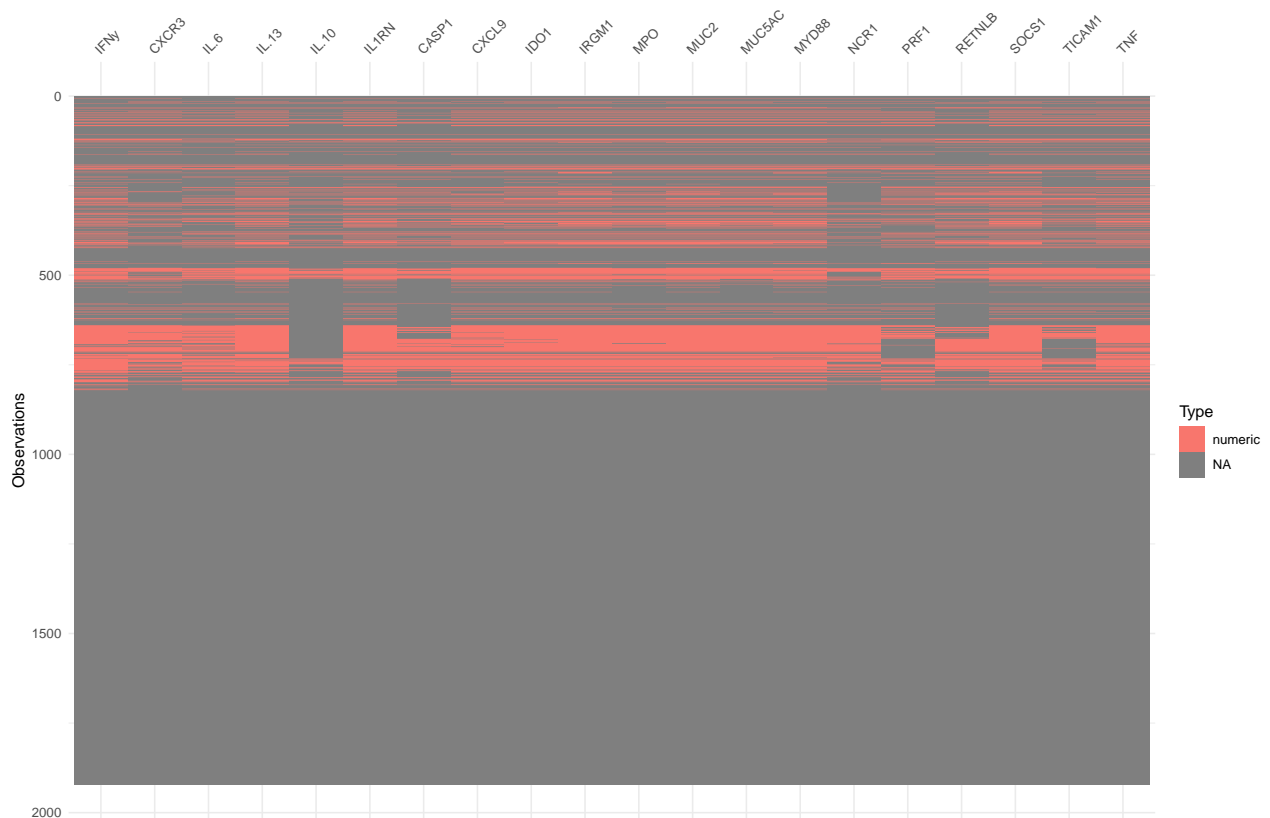
## Genes

```
# field samples
field <- hm %>% filter(origin == "Field")

gf_field <- field %>%
  dplyr::select(all_of(c(Genes_wild)))

vis_dat(gf_field)
```

```
## Warning: `gather()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
```



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

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

#select same rows in the first table
field_genes <- field[row.names(gf_field), ]

#remove wrongly normalized genes
field <- field %>%
```

```

dplyr::select(-ends_with("_N"))

# really removing empty columns
field_genes <- field_genes %>%
  discard(~all(is.na(.) | . == ""))

# looking at patterns of nas
#pattern_na <- as.data.frame(md.pattern(field_genes))

#sapply(field, function(x) sum(is.na(x)))

#select the relevant columns to use for the imputation
field_genes <- field_genes %>%
  dplyr::select(c(Mouse_ID, MC.Eimeria, delta_ct_cewe_MminusE, Sex, Longitude, Latitude,
    Year, mtBamH, YNPARG, X332, X347, X65, Tsx, Btk, Syap1, Es1,
    Gpd1, Idh1, Mpi, Np, Sod1, Es1C, Gpd1C, Idh1C, NpC, Sod1C,
    HI_NLoci, Spleen, Trichuris_muris, Mastophorus_muris,
    Catenotaenia_pusilla, Status,
    Heterakis_sp, N_oocysts_sq1,
    N_oocysts_sq2, N_oocysts_sq3, N_oocysts_sq4, N_oocysts_sq5,
    N_oocysts_sq6, N_oocysts_sq7, N_oocysts_sq8, Region,
    Body_Length, Fleas, Tail_Length, eimeriaSpecies, Ct.Eimeria,
    Ct.Mus, ILWE_Crypto_Ct, Aspiculuris_sp, Syphacia_sp, Taenia_sp,
    Hymenolepis_sp, FEC_Eim_Ct,
    all_of(Genes_wild)))

#had to remove as they were disturbing the imputation: Worms_presence, MC.Eimeria.FEC, Heligmosomoides

#vis_miss(field)

# The frequency distribution of the missing cases per variable can be obtained
# as:
init <- mice(field_genes, maxit = 0)

## Warning: Number of logged events: 24

#we want to impute only the specific variables
meth <- init$method

# m=5 refers to the number of imputed datasets. Five is the default value.
igf <- mice(field_genes, m = 5, seed = 500) # method = meth,

##
## iter imp variable
## 1 1 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 1 2 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 1 3 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 1 4 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 1 5 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 2 1 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 2 2 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 2 3 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3

```

```
## 2 4 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 2 5 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 3 1 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 3 2 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 3 3 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 3 4 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 3 5 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 4 1 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 4 2 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 4 3 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 4 4 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 4 5 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 5 1 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 5 2 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 5 3 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 5 4 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
## 5 5 MC.Eimeria delta_ct_cewe_MminusE HI_NLoci Spleen N_oocysts_sq1 N_oocysts_sq2 N_oocysts_sq3
```

```
## Warning: Number of logged events: 539
```

```
summary(igf)
```

```
## Class: mids
## Number of multiple imputations: 5
## Imputation methods:
##           Mouse_ID      MC.Eimeria delta_ct_cewe_MminusE
##           ""           "logreg"           "pmm"
##           Sex           Longitude           Latitude
##           ""           ""           ""
##           Year           mtBamH           YNPAR
##           ""           ""           ""
##           X332           X347           X65
##           ""           ""           ""
##           Tsx           Btk           Syap1
##           ""           ""           ""
##           Es1           Gpd1           Idh1
##           ""           ""           ""
##           Mpi           Np           Sod1
##           ""           ""           ""
##           Es1C           Gpd1C           Idh1C
##           ""           ""           ""
##           NpC           Sod1C           HI_NLoci
##           ""           ""           "pmm"
##           Spleen      Trichuris_muris      Mastophorus_muris
##           "pmm"           ""           ""
## Catenotaenia_pusilla      Status           Heterakis_sp
##           ""           ""           ""
##           N_oocysts_sq1      N_oocysts_sq2      N_oocysts_sq3
##           "pmm"           "pmm"           "pmm"
##           N_oocysts_sq4      N_oocysts_sq5      N_oocysts_sq6
##           "pmm"           "pmm"           "pmm"
##           N_oocysts_sq7      N_oocysts_sq8      Region
##           "pmm"           "pmm"           ""
##           Body_Length      Fleas           Tail_Length
##           "pmm"           "logreg"           "pmm"
```

```

##      eimeriaSpecies      Ct.Eimeria      Ct.Mus
##      ""                  "pmm"          "pmm"
##      ILWE_Crypto_Ct      Aspicularis_sp  Syphacia_sp
##      "pmm"                ""             ""
##      Taenia_sp           Hymenolepis_sp   FEC_Eim_Ct
##      ""                  ""             "pmm"
##      IFNy                CXCR3           IL.6
##      "pmm"               "pmm"          "pmm"
##      IL.13               IL.10           IL1RN
##      "pmm"               "pmm"          "pmm"
##      CASP1               CXCL9           IDO1
##      "pmm"               "pmm"          "pmm"
##      IRGM1               MPO            MUC2
##      "pmm"               "pmm"          "pmm"
##      MUC5AC              MYD88           NCR1
##      "pmm"               "pmm"          "pmm"
##      PRF1                RETNLB          SOCS1
##      "pmm"               "pmm"          "pmm"
##      TICAM1              TNF
##      "pmm"               "pmm"

## PredictorMatrix:
##      Mouse_ID MC.Eimeria delta_ct_cewe_MminusE Sex Longitude
## Mouse_ID      0      1      1  0      1
## MC.Eimeria    0      0      1  0      1
## delta_ct_cewe_MminusE 0      1      0  0      1
## Sex           0      1      1  0      1
## Longitude     0      1      1  0      0
## Latitude      0      1      1  0      1
##
##      Latitude Year mtBamH YNPARG X332 X347 X65 Tsx Btk Syap1
## Mouse_ID      1      1      0      0      0      0      0      0      0      0
## MC.Eimeria    1      1      0      0      0      0      0      0      0      0
## delta_ct_cewe_MminusE 1      1      0      0      0      0      0      0      0      0
## Sex           1      1      0      0      0      0      0      0      0      0
## Longitude     1      1      0      0      0      0      0      0      0      0
## Latitude      0      1      0      0      0      0      0      0      0      0
##
##      Es1 Gpd1 Idh1 Mpi Np Sod1 Es1C Gpd1C Idh1C NpC Sod1C
## Mouse_ID      0      0      0      0      0      0      0      0      0      0      0
## MC.Eimeria    0      0      0      0      0      0      0      0      0      0      0
## delta_ct_cewe_MminusE 0      0      0      0      0      0      0      0      0      0      0
## Sex           0      0      0      0      0      0      0      0      0      0      0
## Longitude     0      0      0      0      0      0      0      0      0      0      0
## Latitude      0      0      0      0      0      0      0      0      0      0      0
##
##      HI_NLoci Spleen Trichuris_muris Mastophorus_muris
## Mouse_ID      1      1      1      1
## MC.Eimeria    1      1      1      1
## delta_ct_cewe_MminusE 1      1      1      1
## Sex           1      1      1      1
## Longitude     1      1      1      1
## Latitude      1      1      1      1
##
##      Catenotaenia_pusilla Status Heterakis_sp N_oocysts_sq1
## Mouse_ID      1      0      1      1
## MC.Eimeria    1      0      1      1
## delta_ct_cewe_MminusE 1      0      1      1
## Sex           1      0      1      1

```

```

## Longitude          1      0          1          1
## Latitude            1      0          1          1
##                   N_oocysts_sq2 N_oocysts_sq3 N_oocysts_sq4 N_oocysts_sq5
## Mouse_ID           1          1          1          1
## MC.Eimeria         1          1          1          1
## delta_ct_cewe_MminusE 1          1          1          1
## Sex                1          1          1          1
## Longitude          1          1          1          1
## Latitude            1          1          1          1
##                   N_oocysts_sq6 N_oocysts_sq7 N_oocysts_sq8 Region
## Mouse_ID           1          1          1          0
## MC.Eimeria         1          1          1          0
## delta_ct_cewe_MminusE 1          1          1          0
## Sex                1          1          1          0
## Longitude          1          1          1          0
## Latitude            1          1          1          0
##                   Body_Length Fleas Tail_Length eimeriaSpecies Ct.Eimeria
## Mouse_ID           1          1          1          0          1
## MC.Eimeria         1          1          1          0          1
## delta_ct_cewe_MminusE 1          1          1          0          1
## Sex                1          1          1          0          1
## Longitude          1          1          1          0          1
## Latitude            1          1          1          0          1
##                   Ct.Mus ILWE_Crypto_Ct Aspicularis_sp Syphacia_sp
## Mouse_ID           1          1          1          1
## MC.Eimeria         1          1          1          1
## delta_ct_cewe_MminusE 1          1          1          1
## Sex                1          1          1          1
## Longitude          1          1          1          1
## Latitude            1          1          1          1
##                   Taenia_sp Hymenolepis_sp FEC_Eim_Ct IFNy CXCR3 IL.6 IL.13
## Mouse_ID           1          1          1          1          1          1
## MC.Eimeria         1          1          1          1          1          1
## delta_ct_cewe_MminusE 1          1          1          1          1          1
## Sex                1          1          1          1          1          1
## Longitude          1          1          1          1          1          1
## Latitude            1          1          1          1          1          1
##                   IL.10 IL1RN CASP1 CXCL9 IDO1 IRGM1 MPO MUC2 MUC5AC MYD88
## Mouse_ID           1          1          1          1          1          1          1          1          1
## MC.Eimeria         1          1          1          1          1          1          1          1          1
## delta_ct_cewe_MminusE 1          1          1          1          1          1          1          1          1
## Sex                1          1          1          1          1          1          1          1          1
## Longitude          1          1          1          1          1          1          1          1          1
## Latitude            1          1          1          1          1          1          1          1          1
##                   NCR1 PRF1 RETNLB SOCS1 TICAM1 TNF
## Mouse_ID           1          1          1          1          1          1
## MC.Eimeria         1          1          1          1          1          1
## delta_ct_cewe_MminusE 1          1          1          1          1          1
## Sex                1          1          1          1          1          1
## Longitude          1          1          1          1          1          1
## Latitude            1          1          1          1          1          1
## Number of logged events: 539
##   it im dep      meth      out
## 1  0  0      constant Mouse_ID

```

```
## 2 0 0      constant      Sex
## 3 0 0      constant      mtBamH
## 4 0 0      constant      YNPAR
## 5 0 0      constant      X332
## 6 0 0      constant      X347

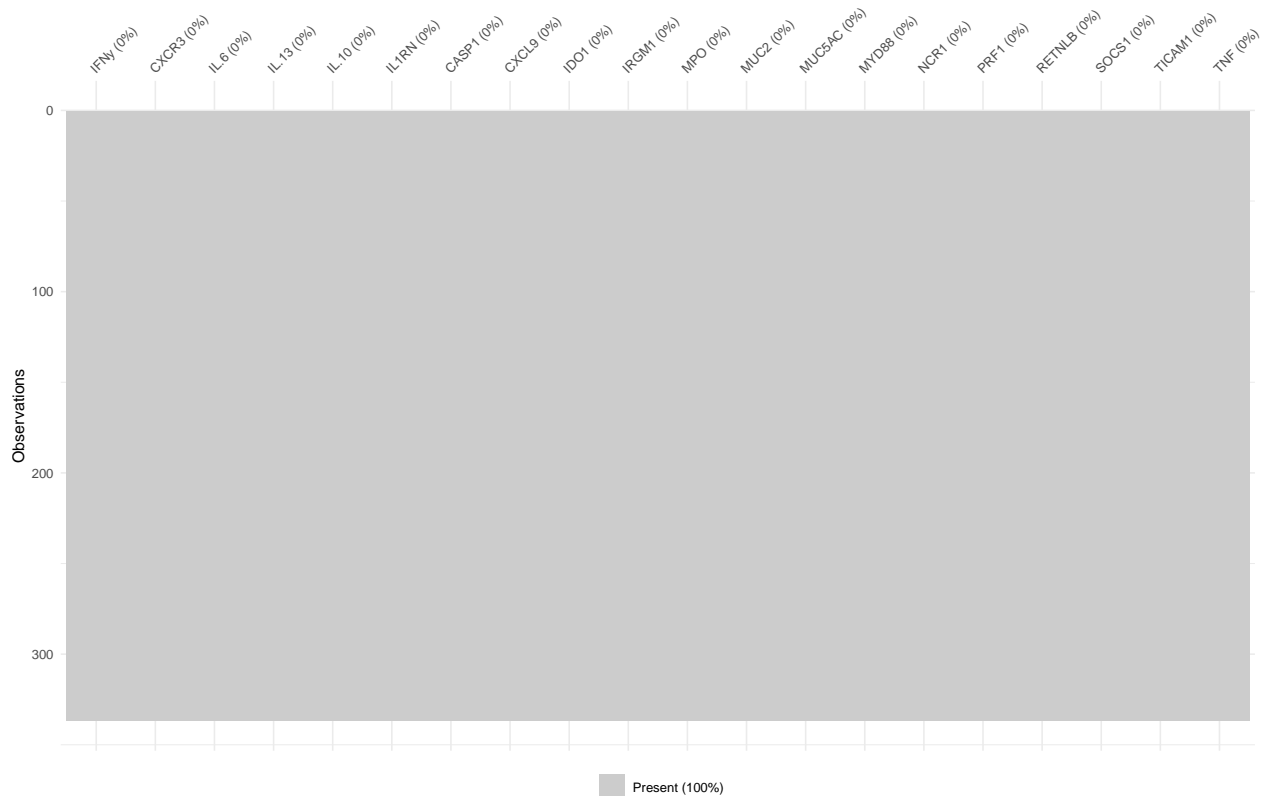
# to check each column with imputed data
## igf$imp$IFNy

#Now we can get back the completed dataset using the complete()
complete_field <- complete(igf, 1)

#sapply(complete_field, function(x) sum(is.na(x)))

# select the required columns
imp_field <- complete_field %>%
  dplyr::select(all_of(Genes_wild))

#visualize missingness
vis_miss(imp_field)
```



```
#add an ending to the imputed columns
colnames(imp_field) <- paste(colnames(imp_field), "imp", sep = "_")

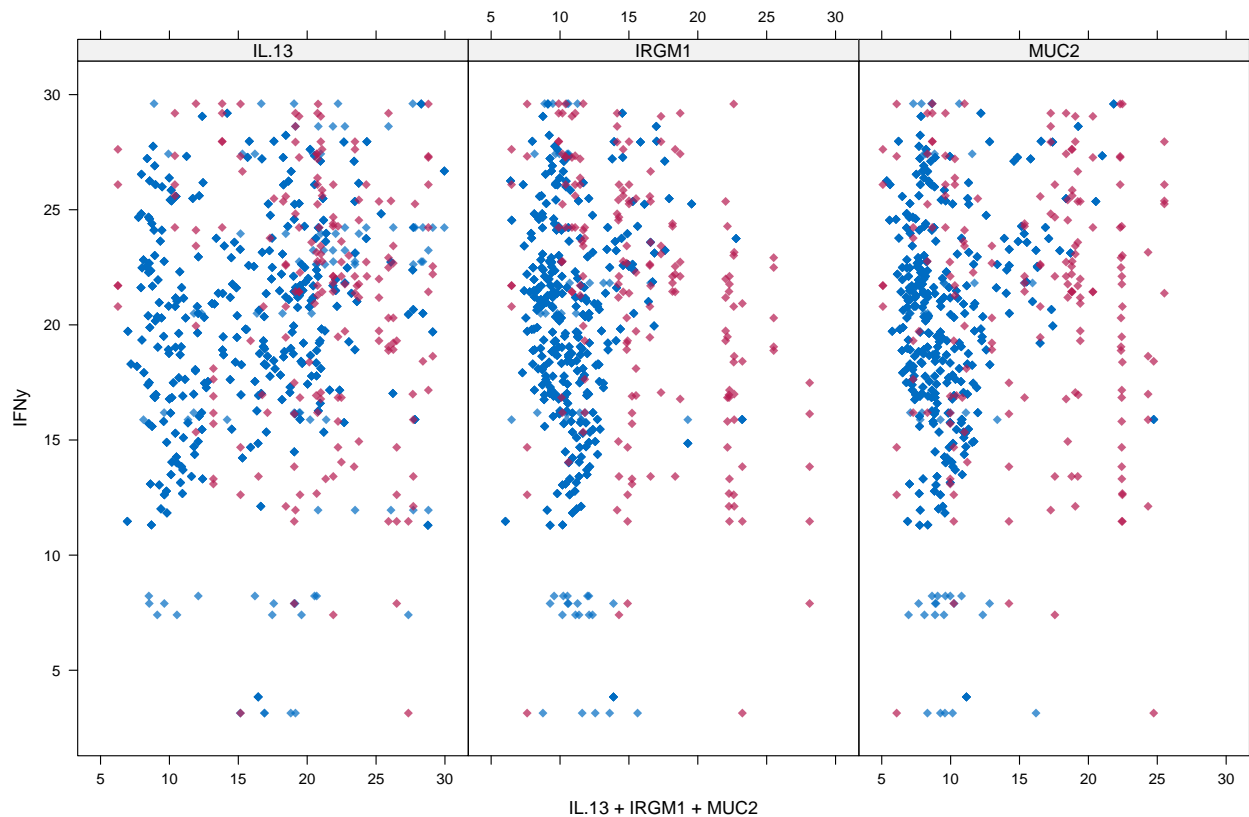
#now join it to the full data set of the laboratory infections
field_genes <- cbind(field_genes, imp_field)
```



```
# join the now imputed columns to the field
field <- field_genes %>%
  left_join(field, by = intersect(colnames(field), colnames(field_genes)))
```

Let's compare the distributions of original and imputed data using a some useful plots. First of all we can use a scatterplot and plot Ozone against all the other variables. Let's first plot the variables for which we have few missing values.

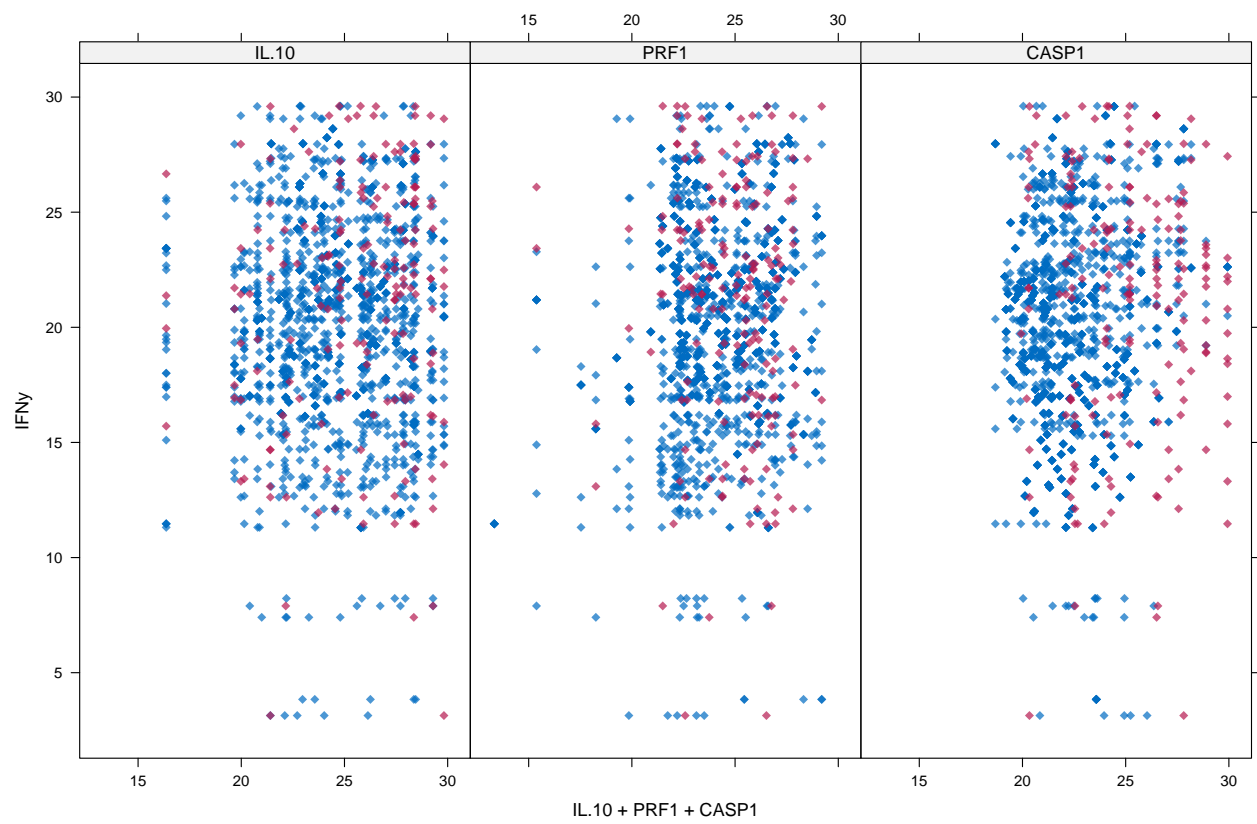
```
xyplot(igf, IFNy ~ IL.13 + IRGM1 + MUC2, pch=18, cex=1)
```



What we would like to see is that the shape of the magenta points (imputed) matches the shape of the blue ones (observed). The matching shape tells us that the imputed values are indeed “plausible values”.

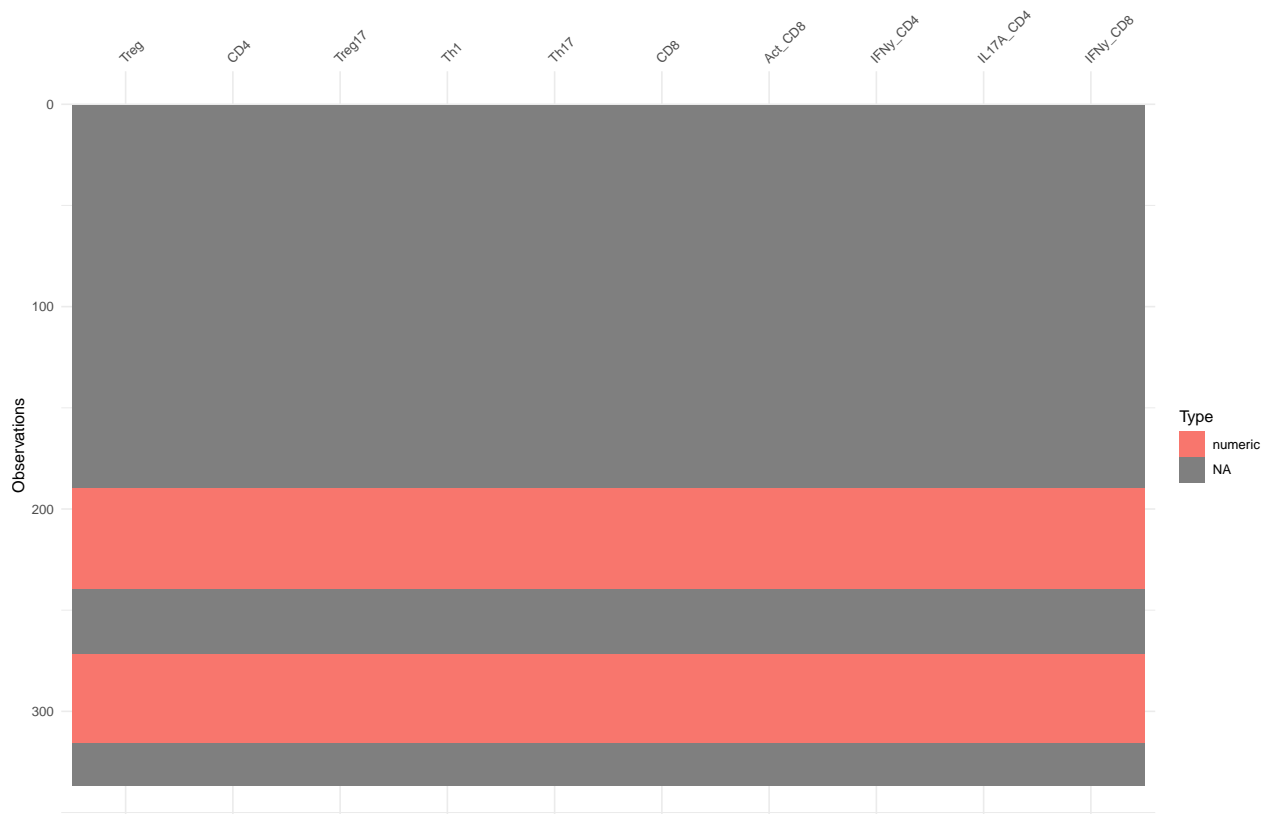
Now let's plot the variables with many missing data points.

```
xyplot(igf, IFNy ~ IL.10 + PRF1 + CASP1, pch=18, cex=1)
```



```
stripplot(igf, pch = 20, cex = 1.2)
```





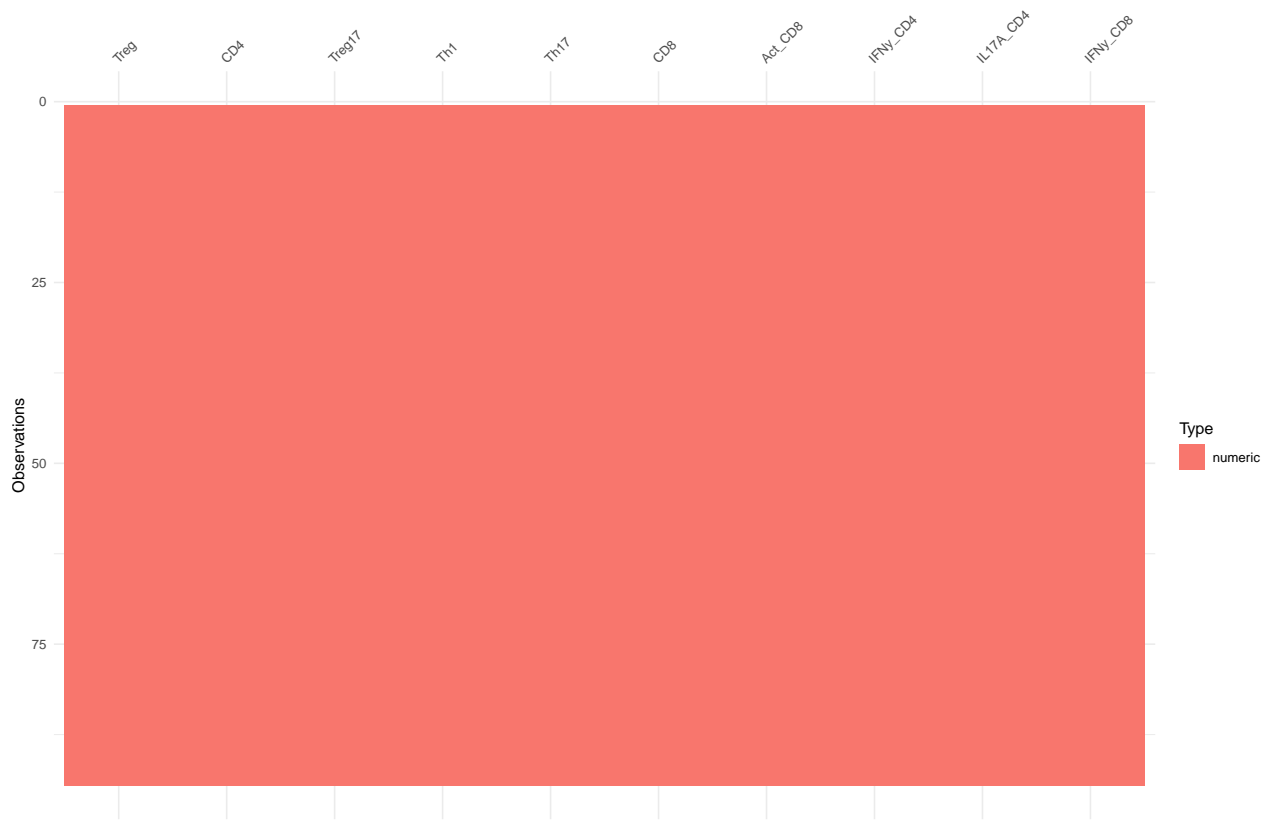
```
#remove rows with only nas
gf_field <- gf_field[,colSums(is.na(gf_field))<nrow(gf_field)]

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

#select same rows in the first table
field_facs <- field[row.names(gf_field), ]

# really removing empty columns
field_facs <- field_facs %>%
  discard(~all(is.na(.) | . == ""))

vis_dat(gf_field)
```



We have no need to impute the facs data, as nothing is missing from the 95 samples

Just add here the lab data and join the two tables

```
lab <- read.csv("output_data/Lab_imputed.csv")

# join the genes and facs data
c <- lab %>%
  full_join(field, by = intersect(colnames(lab), colnames(field)))

##save the imputed data
write.csv(field_genes, "output_data/imputed_MICE.csv", row.names = FALSE)
```