Data Simulation Project

Laura

May 19, 2016

Differential correlation of cytokine responses with age

Project description

This is a data simulation project with the objective to reproduce the cytokine responses described in Hartel, et al. The objective of this study was to characterize age-related immune cell responses, particularly T cell cytokine levels, in order to assess normal variation, state of activation and maturation processes of immunological status in childhood.

Simulating this data has been very complicated, first of all because this study utilizes **nonparametric** methods which do not rely on the estimation of parameters such as the mean or the standard deviation. As a result, I have based my analysis on the correlations (**Spearman's rho**) described in this paper, and also by interpreting all the graphs displayed.

Also, not all the variables described in the paper were measured on the same subjects, as a result, I have divided my data into 5 different tables:

- *Protein production*
  - 5 predictor variables (some correlation among them) and age as a response variable.
- *mRNA expression*
  - 3 predictor variables and age as a response variable (plus age category as factor variable to display the data).
- *Cells producing INFy*
  - 1 predictor variables (correlated with IL12) and age as a response variable.
- *Cells producing TNFa*
  - 1 predictor variables and age as a response variable.
- *Cells producing IL12*
  - 2 predictor variables and age as a response variable.

Data simulation

In order to simulate the correlations described in this paper, I have used Spearman's rho coefficient, which a statistical tool for describing the strength of the monotonic relationship between two ranked variables. I have made use of different functions described in the package GenOrd (Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions) and MASS (to simulate from a multivariate normal distribution), and other previously described functions to generate positive definitive correlation matrixes and marginal distributions. I have made several assumptions related to the distribution of the data in order to generate my random correlated variables. I have assumed a uniform distribution for cytokine production (units: mg/ul) and also for mRNA expression (units: cytokine mRNA expression/106 -actin mRNA copies), and a Poisson distribution for all those variables that represent the percentage of cells expressing cytokines (IL12, INFy, and TNFa).

Here is the code used to simulate cytokine production:

```
N<-1000 # Determine the size of this data set

# Random sample function
ordsamplep<-function (n, lambda, Sigma)
{
  k <- length(lambda)
  valori <- mvrnorm(n, rep(0, k), Sigma)
  for (i in 1:k)
  {
    valori[, i] <- qpois(pnorm(valori[,i]), lambda[i])
  }
  return(valori)
}

# Create a Correlation matrix function

s_mat <- function (n, ev = runif(n, 0, 10))
{
  Z <- matrix(ncol=n, rnorm(n^2))
  decomp <- qr(Z)
  Q <- qr.Q(decomp)
  R <- qr.R(decomp)
  d <- diag(R)
  ph <- d / abs(d)
  O <- Q %*% diag(ph)
  Z <- t(O) %*% diag(ev) %*% O
  return(Z)
}

set.seed(222)
Sigma_m <- s_mat(n=6, ev=3:8)
Sigma_m[2,1]<-1.8 #384
Sigma_m[1,2]<-1.8
Sigma_m[3,1]<-2.2 #414
Sigma_m[1,3]<-2.2
Sigma_m[4,1]<-1.7 #342
Sigma_m[1,4]<-1.7
Sigma_m[3,2]<-3.6 #744
Sigma_m[2,3]<-3.6
Sigma_m[4,2]<-2.4 #579
Sigma_m[2,4]<-2.4
Sigma_m[6,2]<-2.4 #519
Sigma_m[2,6]<-2.4
Sigma_m[4,3]<-3.8 #753
Sigma_m[3,4]<-3.8
Sigma_m[6,3]<-3.5 #736
Sigma_m[3,6]<-3.5
Sigma_m[5,4]<-3.8 #845
Sigma_m[4,5]<-3.8
Sigma_m[6,4]<-3.5 #766
Sigma_m[4,6]<-3.5
Sigma_m[6,5]<-3 #756
```

```
Sigma_m[5,6]<-3
eigen(Sigma_m)$val

set.seed(222)
# Order: age_p, c_IL2, c_TNFa, c_IL4, c_IL5 , c_IL10
lambda <- c(40, 6000, 4000, 300, 1000, 1000)
Sigma<-Sigma_m
df_mg_age <- ordsamplep(N, lambda, Sigma)
colnames(df_mg_age)<-c("age", "IL2_mg", "TNFa_mg", "IL4_mg", "IL5", "IL10")
df_mg_age<-data.frame(df_mg_age)

#Negative values to 0
n_df_mg_age<-df_mg_age
n_df_mg_age$IL2_mg[df_mg_age$IL2_mg<0] <- 0
n_df_mg_age$TNFa_mg[df_mg_age$TNFa_mg<0] <- 0
n_df_mg_age$IL4_mg[df_mg_age$IL4_mg<0] <- 0
```

Here is the code used to simulate mRNA expression:

```
set.seed(222)
Sigma_m <- s_mat(n=4, ev=1:4)
eigen(Sigma_m)$val
# Only IL4 and age are correlated
Sigma_m[4,1]<-0.63
Sigma_m[1,4]<-0.63
eigen(Sigma_m)$val

lambda <- c(12, 5000, 2000,20000)
Sigma<-Sigma_m
df_mRNA_age<- ordsamplep(N, lambda, Sigma)
colnames(df_mRNA_age)<-c("age", "IL2", "TNFa_m", "IL4")
df_mRNA_age<-data.frame(df_mRNA_age)

#Create age category
df_mRNA_age$age_interval<-NA
df_mRNA_age$age_interval[df_mRNA_age$age==0]<-"0 Newborn"
df_mRNA_age$age_interval[df_mRNA_age$age>=1 & df_mRNA_age$age<=6]<-"1-6 months"
df_mRNA_age$age_interval[df_mRNA_age$age>=7 & df_mRNA_age$age<=12]<-"7-12 months"
df_mRNA_age$age_interval[df_mRNA_age$age>=13 & df_mRNA_age$age<=24]<-"13-24 months"
df_mRNA_age$age_interval[df_mRNA_age$age>=25 & df_mRNA_age$age<=48]<-"25-48 months"
df_mRNA_age$age_interval[df_mRNA_age$age>=49 & df_mRNA_age$age<=96]<-"49-96 months"
```

Here is the code used to simulate INFy and TNFa expression:

```
set.seed(222)
age_IFNy<-c(rep(120, N*(0.14)), rep(0,N*(0.2)), rep(-20,N*(0.23)), ceiling(runif(N*(0.43), 1, 96)))
age_df <- as.data.frame(replicate(1, sample(age_IFNy, N, replace = F)))

sampling <- function(n, rho, X1) {

  C <- matrix(rho, nrow = 2, ncol = 2)
  diag(C) <- 1
  C <- chol(C)
```

```
  X2 <- runif(N, 0, 60)
  X <- cbind(X1,X2)
  # Induce correlation
  df <- X %*% C
  return(df)
}

df_IFNy_age<-sampling(N,0.4, age_df[,1])
colnames(df_IFNy_age)<-c("age", "IFNy")
df_IFNy_age<-data.frame(df_IFNy_age)

#Negative values to 0
n_df_IFNy_age<-df_IFNy_age
n_df_IFNy_age$IFNy[n_df_IFNy_age$IFNy<0] <- 0

#Age categories
n_df_IFNy_age$category[n_df_IFNy_age$age==-20]<-'preterm'
n_df_IFNy_age$category[n_df_IFNy_age$age== 0]<-"newborn"
n_df_IFNy_age$category[n_df_IFNy_age$age>0 & n_df_IFNy_age$age<=96]<-"Infant"
n_df_IFNy_age$category[n_df_IFNy_age$age== 120]<-"Adult"

#############################################################################

set.seed(222)
age_TNFa<-c(rep(120, N*(0.19)), rep(0,N*(.13)), rep(-20,N*(.14)), ceiling(runif(N*(.54), 1, 96)))
age_df <- as.data.frame(replicate(1, sample(age_TNFa, N, replace = F)))

sampling <- function(n, rho, X1) {
  C <- matrix(rho, nrow = 2, ncol = 2)
  diag(C) <- 1
  C <- chol(C)
  X2 <- runif(N, 0, 40)
  X <- cbind(X1,X2)
  # Induce correlation
  df <- X %*% C
  return(df)
}

df_TNFa_age<-sampling(N,0.315, age_df[,1])
colnames(df_TNFa_age)<-c("age", "TNFa")
df_TNFa_age<-data.frame(df_TNFa_age)

#Negative values to 0
n_df_TNFa_age<-df_TNFa_age
n_df_TNFa_age$TNFa[n_df_TNFa_age$TNFa<0] <- 0

#Age category
n_df_TNFa_age$category[n_df_TNFa_age$age==-20]<-'preterm'
n_df_TNFa_age$category[n_df_TNFa_age$age== 0]<-"newborn"
n_df_TNFa_age$category[n_df_TNFa_age$age>0 & n_df_TNFa_age$age<=96]<-"Infant"
n_df_TNFa_age$category[n_df_TNFa_age$age== 120]<-"Adult"
```

Here is the code used to simulate IL12 producing cells:

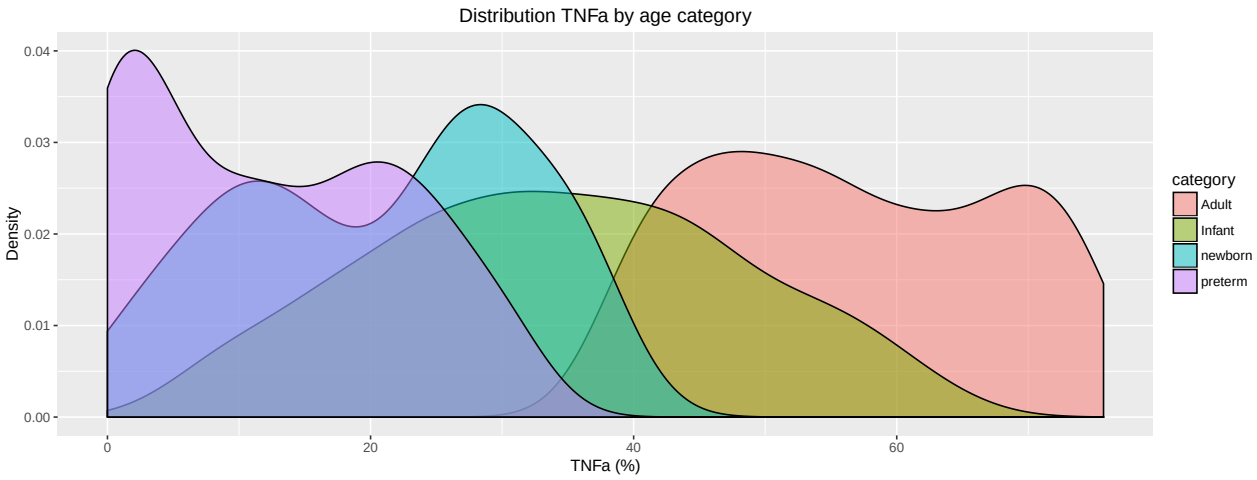
```
set.seed(222)
v_IFNy<-df_IFNy_age[,2]
age_IL12<-c(rep(0,N*(.16)), ceiling(runif(N*(0.84), 1, 96)))
age_df <- as.data.frame(replicate(1, sample(age_IL12, N, replace = F)))

sampling <- function(n, rho, X1) {
  C <- matrix(rho, nrow = 2, ncol = 2)
  diag(C) <- 1
  C <- chol(C)
  X2 <- runif(N, 0, 50)
  X <- cbind(X1,X2)
  # Induce correlation
  df <- X %*% C
  return(df)
}

df_IL12_age<-sampling(N,0.15, age_df[,1])
colnames(df_IL12_age)<-c("age", "IL12")
df_IL12_age<-data.frame(df_IL12_age)
#Add a column for TFGb
df_IL12_age$TFGb<-c(runif(N, 0, 50))

#Negative values to 0
n_df_IL12_age<-df_IL12_age
n_df_IL12_age$IL12[n_df_IL12_age$IL12<0] <- 0
```

Data analysis

Distribution of Cytokine expression

Once the data set was created is important to verify the distribution of the data, which as expected will have uniform distributions or Poisson (which behave as normally distributed because of the high number of sample values randomly generated).

Figure 1



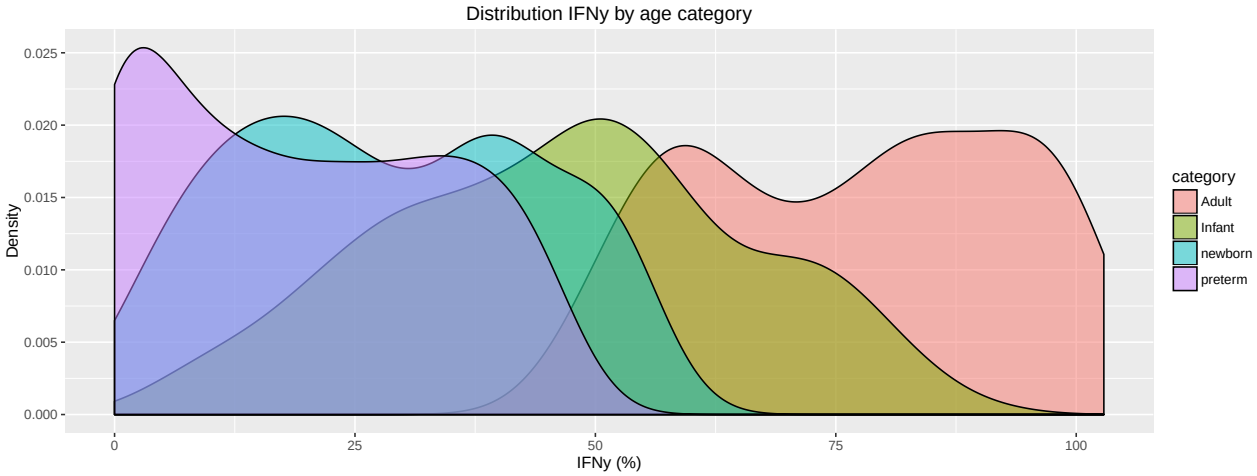


Figure 1 allows to clearly identify a correlation with age of TNFa and IFNy. The mean percentage of cells expressing these cytokines increases as age increases, a lower mean expression is found in preterm babies and the highest is of course in adults.

Figure 2

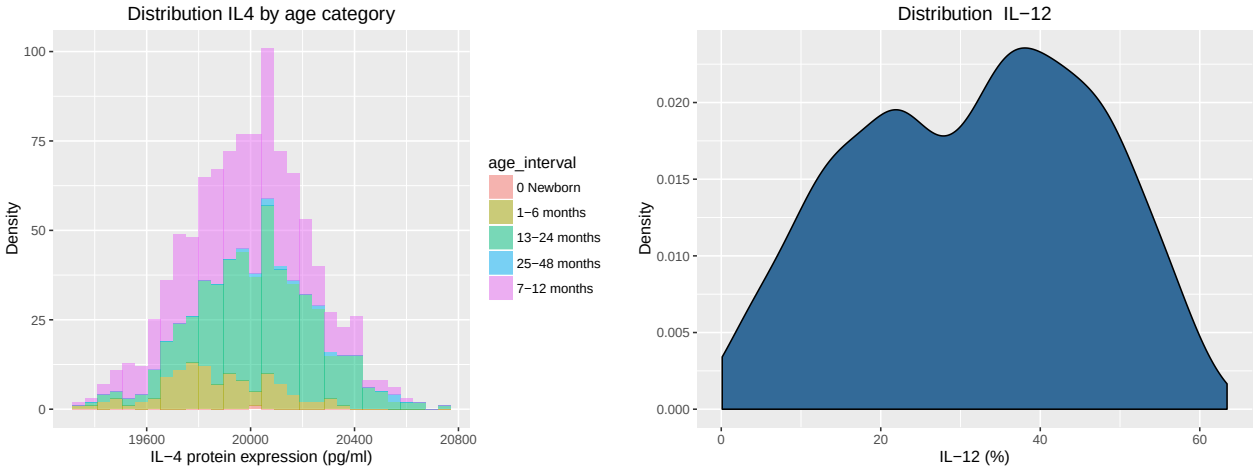


Figure 2 shows how IL4 is almost normally distributed, and its expression is pretty much the same in every age interval, where most of the data is represented by 1-year-old babies.

The percentage of cells expressing IL12, on the other hand, show a uniform distribution (only children from 1-96 months is represented in this graph).

Figure 3

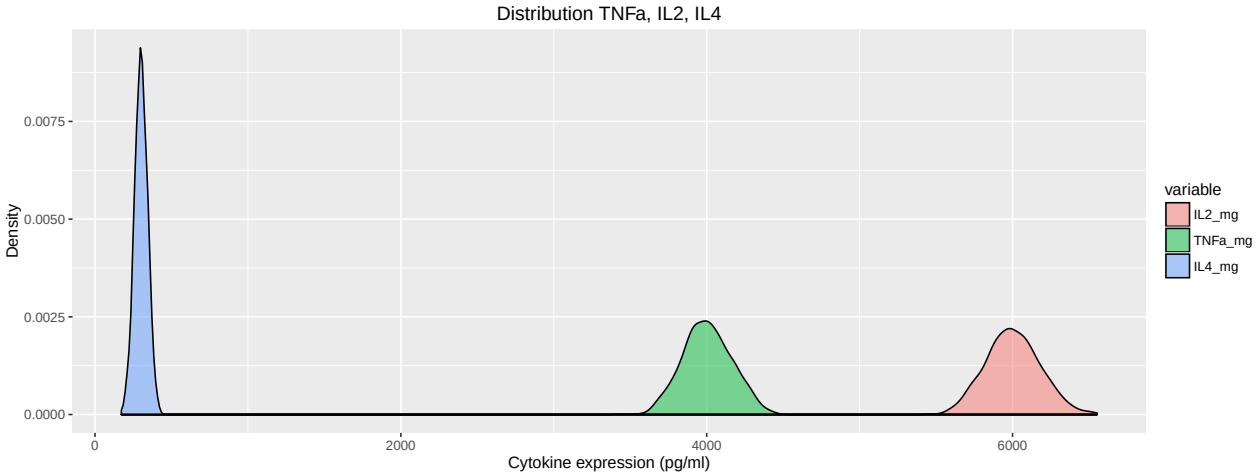


Figure 3 demonstrates how distinct is the proportion of intracellular expression of IL2, TNFa, and IL4. Where the highest is interleukin-2 an important inflammatory cytokine.

Correlations

And here are some of the `Sparman's correlation test` that demostrate I was able to reproduce succesfully those variable reponses.

IFN y

```
##
##  Spearman's rank correlation rho
##
## data:  df_IFNy_age[, 1] and df_IFNy_age[, 2]
## S = 40931000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7544168
```

TNF a

```
##
##  Spearman's rank correlation rho
##
## data:  df_TNFa_age[, 1] and df_TNFa_age[, 2]
## S = 31625000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8102471
```

Protein expression

```
## IL2
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_mg_age[, 1] and df_mg_age[, 2]
## S = 102540000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3847733
```

```
## TNF a
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_mg_age[, 1] and df_mg_age[, 3]
## S = 100510000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3969537
```

```
## IL4
```

```
##
##  Spearman's rank correlation rho
##
## data:  df_mg_age[, 1] and df_mg_age[, 4]
## S = 114530000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3127928
```

IL12

```
##
##  Spearman's rank correlation rho
##
## data:  df_IL12_age[, 1] and df_IL12_age[, 2]
## S = 116640000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.300161
```

RNA expression

```
## IL4
```

```
##
##  Spearman's rank correlation rho
```

```
##
## data:  df_mRNA_age[, 1] and df_mRNA_age[, 4]
## S = 122400000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.265575
```

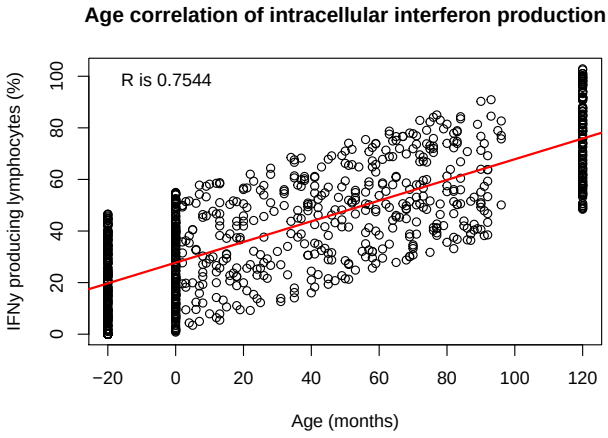
According to the Spearman correlation test, every one of this variables has a significant correlation with age (I was able to reproduce very closely the correlations described in the paper. *IFNy: 0.748, TNFa: 0.784, IL2: 0.384, TNFa: 0.414, IL4: 0.342, IL12:0.331, IL4: 0.29*), where the highest positive correlation found was tumour necrosis factor-alpha (TNF-alpha) an inflammatory cytokine characteristic of T helper type 1 (Th1) cells.

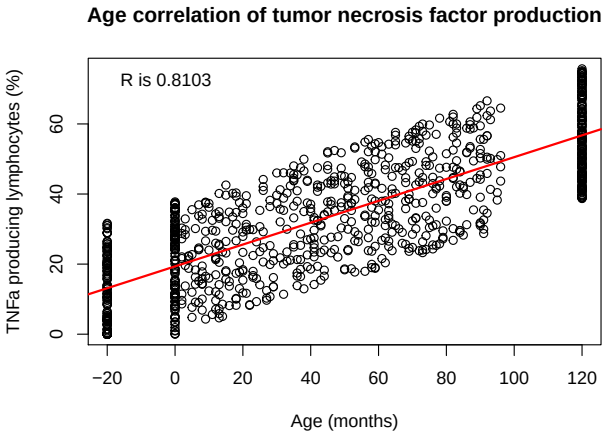
Other correlations found (and also described in the paper) were:

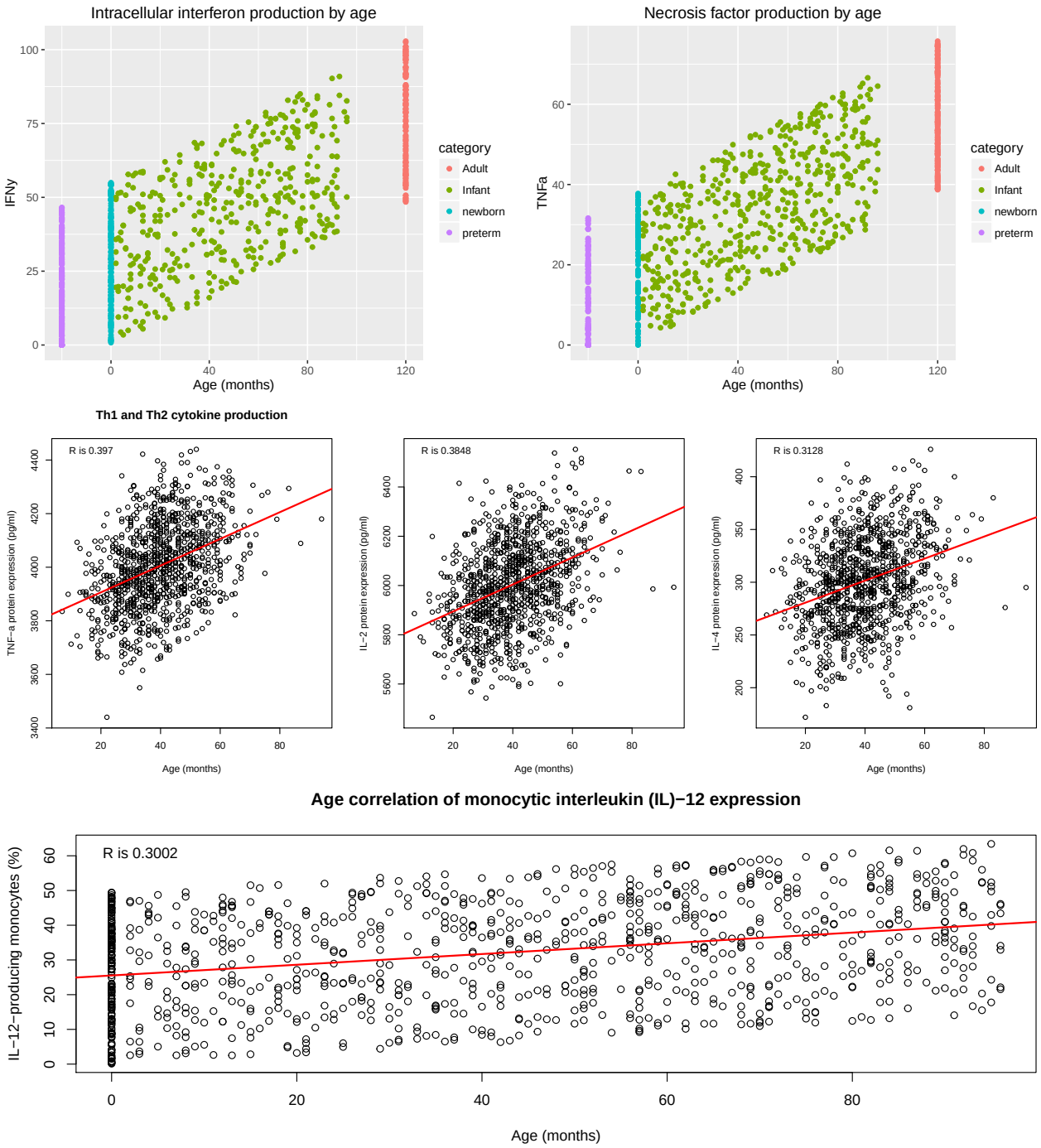
- IL-2 and TNFa : 0.5932552,
- IL-2 and IL-4: 0.4027042,
- IL-2 and IL-10: 0.3916589,
- TNFa and IL-4: 0.5686302,
- TNFa and IL-10: 0.5438225,
- IL-4 and IL-5: 0.7346752,
- IL-4 and IL-10: 0.5542395, and
- IL-5 and IL-10: 0.5328616.

Plots

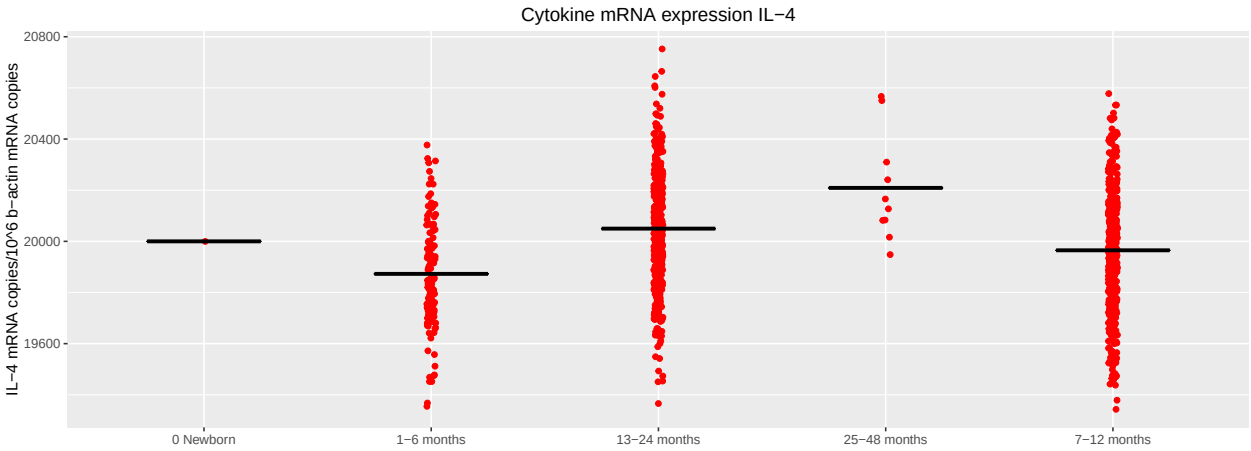
I can also demonstrate the most highly correlated cytokine responses with the following plots:







This plot, on the other hand, shows the distribution by age intervals of IL4 mRNA expression.



Predicting age by cytokine expression

Based on the protein expression levels age can be predicted, as demostrated by the following model:

```
##
## Call:
## lm(formula = age ~ TNFa_mg + IL10 + IL5 + IL4_mg + IL2_mg, data = n1_df_mg_age)
##
## Coefficients:
## (Intercept)      TNFa_mg         IL10          IL5       IL4_mg
##  -993.48836      0.22996     -0.31827      0.60001     -0.81200
##      IL2_mg
##     0.01272

##
## Call:
## lm(formula = age ~ ., data = n1_df_mg_age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6076 -1.4768 -0.2262  1.4443 10.9882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.935e+02  7.375e+00 -134.70   <2e-16 ***
## IL2_mg       1.272e-02  5.400e-04   23.55   <2e-16 ***
## TNFa_mg      2.300e-01  1.738e-03  132.31   <2e-16 ***
## IL4_mg      -8.120e-01  7.588e-03 -107.01   <2e-16 ***
## IL5          6.000e-01  4.745e-03  126.46   <2e-16 ***
## IL10        -3.183e-01  2.181e-03 -145.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.373 on 993 degrees of freedom
## Multiple R-squared:  0.9644, Adjusted R-squared:  0.9642
## F-statistic:  5384 on 5 and 993 DF,  p-value: < 2.2e-16
```

In order to demonstrate how strong age correlates with cytokine responses as the immune system matures, I have created a linear regression model where all the variability of the response variable “age” can be

explained by the intracellular expression (mg/ul) of five important predictor variables (cytokines expressed by T helper cells required for host defense).

Both, the model and all the variables are significant (p-value < 0.05) and explain most of the variability of the response variable (very high R-squared).

The equation of the fitted model can be summarized as follows:

$$AGE = -993.48836 + 0.22996(TNFa) - 0.31827(IL10) + 0.6001(IL5) - 0.81200(IL4) + 0.01272(IL2)$$

Cytokine profile of TH1 and TH2

Finally, I am presenting a summary of the cytokine profiles of TH1 and TH2 cells, which is important to establish a baseline for T helper cytokine levels of a normal immune state in childhood.

```
## Th1
```

```
##        IL2     TNFa     INFy
## 1 6002.971 4004.523 41.02227
```

```
## Th2
```

```
##    IL4_mg       IL5      IL10
##  301.1532  999.7778 1003.3844
```

Conclusions

I have successfully generated a data set that reproduced a cross-sectional analysis and observational study where different cytokine responses were analyzed at different age stages.

This has been a very complicated process due to the limited information available about the experimental data produced in this investigation, and the nature of the nonparametric analysis.

However, this type of simulation represents a very important resource to help formulate hypotheses on the normal 'ontogeny of immune cells from birth to childhood' because is very difficult to obtain infant's blood samples. And of course, the outcome of this type of analysis can be further utilized as a reference value of cytokine production to diagnose and monitor immune-mediated disorders in young children Hartel, et al.

References

Here are all the packages and functions used to create this project.

Cross validated

Positive Definite matrix

Package GenOrd

Package Mass

Figures side by side

Plot R value

Mean segments