Capstone project ~ Data Wrangling

Elena Tortosa

2/1/2019

Libraries and work directory

I loaded all required libraries and set the work directory.

```
library(tidyr)
library(dplyr)
setwd("/Users/Tortosae/Desktop/Data science course/Capstone_project")
```

Data tables

All data was provided in 14 different excel sheets. I saved all as .csv and loaded them in R. They were named with a number (df#) followed by the name of the perturbation + treatment each table contains.

```
df1_cd3cd28 <- read.table(file="1.cd3cd28.csv",sep=",", header=TRUE)
df2_cd3cd28icam2 <- read.table(file="2. cd3cd28icam2.csv",sep=",", header=TRUE)
df3_cd3cd28aktinhib <- read.table(file="3. cd3cd28aktinhib.csv",sep=",", header=TRUE)
df4_cd3cd28g0076 <- read.table(file="4. cd3cd28g0076.csv",sep=",", header=TRUE)
df5_cd3cd28psitect <- read.table(file="5. cd3cd28psitect.csv",sep=",", header=TRUE)
df6_cd3cd28u0126 <- read.table(file="6. cd3cd28u0126.csv",sep=",", header=TRUE)
df7_cd3cd28ly <- read.table(file="7. cd3cd28ly.csv",sep=",", header=TRUE)
df8_pma <- read.table(file="8. pma.csv",sep=",", header=TRUE)
df9_b2camp <- read.table(file="9. b2camp.csv",sep=",", header=TRUE)
df10_cd3cd28icam2aktinhib <- read.table(file="10. cd3cd28icam2aktinhib.csv",sep=",", header=TRUE)
df11_cd3cd28icam2g0076 <- read.table(file="11. cd3cd28icam2g0076.csv",sep=",", header=TRUE)
df12_cd3cd28icam2psit <- read.table(file="12. cd3cd28icam2psit.csv",sep=",", header=TRUE)
df13_cd3cd28icam2u0126 <- read.table(file="13. cd3cd28icam2u0126.csv",sep=",", header=TRUE)
df14_cd3cd28icam2ly <- read.table(file="14. cd3cd28icam2ly.csv",sep=",", header=TRUE)
```

Column names

Only df8 contained two columns with different names (in lower case). I unfied column names.

```
df8_pma <- df8_pma %>% rename (PIP2 = pip2, PIP3 = pip3)
```

New column for perturbations

Measurments are obtained from two different perturbations ("general perturbation" : GP1 and GP2). At the same time, these perturbations are combined with different treatments (considered below). I added a new column called GP to each table to classify the data depending on the general perturbation is applied : GP = 1 for GP1 and GP = 2 for GP2.

```
df1_cd3cd28 <- df1_cd3cd28 %>%
              mutate(treatment = "cd3cd28") %>% mutate (GP = 1)
df2_cd3cd28icam2 <- df2_cd3cd28icam2 %>%
                    mutate(treatment = "cd3cd28icam2") %>% mutate (GP = 2)
```

```
df3_cd3cd28aktinhib <- df3_cd3cd28aktinhib %>%
                       mutate(treatment = "cd3cd28aktinhib") %>% mutate (GP = 1)
df4_cd3cd28g0076 <- df4_cd3cd28g0076 %>%
                    mutate(treatment = "cd3cd28g0076") %>% mutate (GP = 1)
df5_cd3cd28psitect <- df5_cd3cd28psitect %>%
                       mutate(treatment = "cd3cd28psitect") %>% mutate (GP = 1)
df6_cd3cd28u0126 <- df6_cd3cd28u0126 %>%
                    mutate(treatment = "cd3cd28u0126") %>% mutate (GP = 1)
df7_cd3cd28ly <- df7_cd3cd28ly %>%
                  mutate(treatment = "cd3cd28ly") %>% mutate (GP = 1)
df8_pma <- df8_pma %>%
            mutate(treatment = "pma") %>% mutate (GP = 1)
df9_b2camp <- df9_b2camp %>%
              mutate(treatment = "b2camp") %>% mutate (GP = 1)
df10_cd3cd28icam2aktinhib <-df10_cd3cd28icam2aktinhib %>%
                            mutate(treatment = "cd3cd28icam2aktinhib") %>% mutate (GP = 2)
df11_cd3cd28icam2g0076 <-df11_cd3cd28icam2g0076 %>%
                          mutate(treatment = "cd3cd28icam2g0076") %>% mutate (GP = 2)
df12_cd3cd28icam2psit <- df12_cd3cd28icam2psit %>%
                         mutate(treatment = "cd3cd28icam2psit") %>% mutate (GP = 2)
df13_cd3cd28icam2u0126 <- df13_cd3cd28icam2u0126 %>%
                          mutate(treatment = "cd3cd28icam2u0126") %>% mutate (GP = 2)
df14_cd3cd28icam2ly <- df14_cd3cd28icam2ly %>%
                         mutate(treatment = "cd3cd28icam2ly") %>% mutate (GP = 2)
```

Unique table

I created a unique table for all the perturbations/treatments.

```
alldf <- bind_rows(df1_cd3cd28, df2_cd3cd28icam2, df3_cd3cd28aktinhib, df4_cd3cd28g0076,     df5_cd3cd
```

Dummy variables for the different treatments

I created dummy variables for each of the treatments (independently of the perturbation)

```
alldf <- alldf %>%
          mutate (Akt_inh1 = ifelse (treatment == c("cd3cd28aktinhib", "cd3cd28icam2aktinhib"), 1, 0)) %
          mutate (PKC_inh = ifelse (treatment == c("cd3cd28g0076", "cd3cd28icam2g0076"), 1, 0)) %>%
          mutate (PIP2_inh = ifelse (treatment == c("cd3cd28psitect", "cd3cd28icam2psit"), 1, 0)) %>%
          mutate (MEK_inh = ifelse (treatment == c("cd3cd28u0126", "cd3cd28icam2u0126"), 1, 0)) %>%
          mutate (Akt_inh2 = ifelse (treatment == c("cd3cd28ly", "cd3cd28icam2ly"), 1, 0)) %>%
          mutate (PKC_act = ifelse (treatment == "pma", 1, 0)) %>%
          mutate (PKA_act = ifelse (treatment == "b2camp", 1, 0))
```

Reorder columns

I reorder the columns to have the treatment names and dummy variables first, and after that all the measurments done.

```
alldf <- alldf %>% select(treatment, GP, Akt_inh1, PKC_inh, PIP2_inh, MEK_inh, Akt_inh2, PKC_act, PKA_a
```

Save table

I saved the tabble as .csv document (capstone_project.csv)

```
write.table(alldf, file = "capstone_project.csv", sep = ",", col.names = NA)
```