processing\_script

Amanda Skarlupka

10/26/2019

load the packages needed

library(readxl)  
library(knitr)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2   
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3   
## ✔ tidyr 1.0.0.9000 ✔ stringr 1.4.0   
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl)  
library(gtools)  
library(gdata)

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##   
## Attaching package: 'gdata'

## The following objects are masked from 'package:dplyr':  
##   
## combine, first, last

## The following object is masked from 'package:purrr':  
##   
## keep

## The following object is masked from 'package:stats':  
##   
## nobs

## The following object is masked from 'package:utils':  
##   
## object.size

## The following object is masked from 'package:base':  
##   
## startsWith

load the data. The data is contained within one excel file, but is separated on different sheets

p1\_mab\_full <- read\_excel("~/Documents/Data Science/AmandaSkarlupka-Project/data/raw\_data/data.xlsx",   
 sheet = "p1\_mab", na = "NA")  
ca09\_mab\_full <- read\_excel("~/Documents/Data Science/AmandaSkarlupka-Project/data/raw\_data/data.xlsx",   
 sheet = "ca09\_mab", na = "NA")  
p1\_sera\_full <- read\_excel("~/Documents/Data Science/AmandaSkarlupka-Project/data/raw\_data/data.xlsx",   
 sheet = "p1\_sera", na = "NA")  
ca09\_sera\_full <- read\_excel("~/Documents/Data Science/AmandaSkarlupka-Project/data/raw\_data/data.xlsx",   
 sheet = "ca09\_sera", na = "NA")  
antigen\_key <- read\_excel("~/Documents/Data Science/AmandaSkarlupka-Project/data/raw\_data/data.xlsx",   
 sheet = "antigen\_key", na = "NA")  
antibody\_key <- read\_excel("~/Documents/Data Science/AmandaSkarlupka-Project/data/raw\_data/data.xlsx",  
 sheet = "antibody\_key", na = "NA")

#Take a look at the data and make sure that it loaded correctly

summary(p1\_mab\_full)

## clone\_name Spain/2003 Zhejiang/07 Swine/31   
## Length:12 Min. : 0.0780 Min. : 0.0780 Min. : 1.25   
## Class :character 1st Qu.: 0.2145 1st Qu.: 0.9765 1st Qu.:20.00   
## Mode :character Median : 0.9375 Median : 1.8750 Median :20.00   
## Mean : 7.5260 Mean : 8.8834 Mean :16.88   
## 3rd Qu.:20.0000 3rd Qu.:20.0000 3rd Qu.:20.00   
## Max. :20.0000 Max. :20.0000 Max. :20.00   
## Illinois/09 Minnesota/09 Nebraska/13 Iowa/73   
## Min. :20 Min. : 0.1562 Min. : 1.25 Min. : 0.1562   
## 1st Qu.:20 1st Qu.: 0.5469 1st Qu.: 8.75 1st Qu.:17.5000   
## Median :20 Median :15.0000 Median :20.00 Median :20.0000   
## Mean :20 Mean :11.0547 Mean :15.10 Mean :16.4714   
## 3rd Qu.:20 3rd Qu.:20.0000 3rd Qu.:20.00 3rd Qu.:20.0000   
## Max. :20 Max. :20.0000 Max. :20.00 Max. :20.0000   
## WI/97 Colorado/09 NC/34543/09 MN/15   
## Min. : 0.3125 Min. : 0.1560 Min. : 0.1560 Min. : 2.50   
## 1st Qu.: 1.2500 1st Qu.: 0.5859 1st Qu.: 0.5859 1st Qu.:20.00   
## Median :10.0000 Median : 2.5000 Median : 3.7500 Median :20.00   
## Mean :10.4427 Mean : 8.1510 Mean : 6.6927 Mean :17.29   
## 3rd Qu.:20.0000 3rd Qu.:20.0000 3rd Qu.:10.0000 3rd Qu.:20.00   
## Max. :20.0000 Max. :20.0000 Max. :20.0000 Max. :20.00   
## Utah/09 NC/09 Missouri/13 NC/15 Indiana/00   
## Min. :20 Min. : 0.078 Min. :20 Min. :20 Min. : 0.234   
## 1st Qu.:20 1st Qu.: 1.062 1st Qu.:20 1st Qu.:20 1st Qu.:20.000   
## Median :20 Median : 6.250 Median :20 Median :20 Median :20.000   
## Mean :20 Mean : 9.762 Mean :20 Mean :20 Mean :18.353   
## 3rd Qu.:20 3rd Qu.:20.000 3rd Qu.:20 3rd Qu.:20 3rd Qu.:20.000   
## Max. :20 Max. :20.000 Max. :20 Max. :20 Max. :20.000   
## NC/01 NC/5043-1/09   
## Min. :20 Min. : 0.156   
## 1st Qu.:20 1st Qu.: 2.188   
## Median :20 Median :20.000   
## Mean :20 Mean :12.213   
## 3rd Qu.:20 3rd Qu.:20.000   
## Max. :20 Max. :20.000

str(p1\_mab\_full)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 12 obs. of 19 variables:  
## $ clone\_name : chr "1F8" "3H6" "3D3" "2A5" ...  
## $ Spain/2003 : num 7.5 1.25 0.234 0.078 0.156 ...  
## $ Zhejiang/07 : num 20 1.25 2.5 0.078 0.156 20 20 0.117 1.25 1.25 ...  
## $ Swine/31 : num 20 20 20 1.25 20 1.25 20 20 20 20 ...  
## $ Illinois/09 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ Minnesota/09: num 20 0.312 20 1.25 0.312 ...  
## $ Nebraska/13 : num 20 20 5 1.25 5 20 20 10 20 20 ...  
## $ Iowa/73 : num 20 0.156 20 20 20 ...  
## $ WI/97 : num 20 1.875 10 1.25 0.625 ...  
## $ Colorado/09 : num 20 0.625 2.5 1.25 0.469 ...  
## $ NC/34543/09 : num 1.25 0.469 5 2.5 0.312 ...  
## $ MN/15 : num 20 5 20 20 20 20 20 20 20 2.5 ...  
## $ Utah/09 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ NC/09 : num 0.5 20 1.25 0.078 0.312 ...  
## $ Missouri/13 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ NC/15 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ Indiana/00 : num 0.234 20 20 20 20 20 20 20 20 20 ...  
## $ NC/01 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ NC/5043-1/09: num 20 2.5 20 20 1.25 20 2.5 0.156 20 0.156 ...

str(ca09\_mab\_full)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 18 obs. of 19 variables:  
## $ clone\_name : chr "1E6" "2A12" "5B\_2A12" "2B11" ...  
## $ Spain/2003 : num 0.156 20 0.039 20 20 20 10 2.5 20 0.039 ...  
## $ Zhejiang/07 : num 0.156 20 0.078 20 20 20 20 5 20 0.039 ...  
## $ Swine/31 : num 5 20 20 20 20 2.5 20 0.625 20 20 ...  
## $ Illinois/09 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ Minnesota/09: num 0.312 0.078 20 20 20 ...  
## $ Nebraska/13 : num 0.625 20 20 20 20 5 20 2.5 20 5 ...  
## $ Iowa/73 : num 20 20 20 20 20 20 1.25 5 20 0.156 ...  
## $ WI/97 : num 0.625 20 0.625 20 20 ...  
## $ Colorado/09 : num 0.469 20 0.156 10 20 5 1.25 1.25 20 0.234 ...  
## $ NC/34543/09 : num 0.312 5 0.156 5 20 ...  
## $ MN/15 : num 20 20 0.156 20 20 ...  
## $ Utah/09 : num 20 10 20 20 20 20 20 20 20 20 ...  
## $ NC/09 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ Missouri/13 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ NC/15 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ Indiana/00 : num 5 20 20 20 20 10 20 0.078 20 20 ...  
## $ NC/01 : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ NC/5043-1/09: num 0.078 20 0.02 20 20 ...

str(p1\_sera\_full)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 10 obs. of 19 variables:  
## $ mouse : num 1 2 3 4 5 6 7 8 9 10  
## $ Spain/2003 : num 320 NA 160 320 640 320 320 640 960 NA  
## $ Zhejiang/07 : num 320 NA 80 160 320 160 160 640 320 NA  
## $ Swine/31 : logi NA NA NA NA NA NA ...  
## $ Illinois/09 : logi NA NA NA NA NA NA ...  
## $ Minnesota/09: logi NA NA NA NA NA NA ...  
## $ Nebraska/13 : logi NA NA NA NA NA NA ...  
## $ Iowa/73 : num 320 480 5 640 240 320 5 640 640 160  
## $ WI/97 : num 240 320 160 320 480 240 80 160 400 80  
## $ Colorado/09 : num 960 960 160 320 960 480 320 960 480 320  
## $ NC/34543/09 : num 480 960 160 320 960 320 320 1280 480 320  
## $ MN/15 : num 5 80 5 5 5 5 5 5 5 5  
## $ Utah/09 : logi NA NA NA NA NA NA ...  
## $ NC/09 : logi NA NA NA NA NA NA ...  
## $ Missouri/13 : num 5 NA 5 5 5 5 5 5 5 NA  
## $ NC/15 : logi NA NA NA NA NA NA ...  
## $ Indiana/00 : num 5 5 5 5 5 5 5 5 5 5  
## $ NC/01 : num 5 5 5 5 5 5 5 5 5 5  
## $ NC/5043-1/09: num 1280 2560 5 320 1280 5 640 2560 5 640

str(ca09\_sera\_full)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 6 obs. of 19 variables:  
## $ mouse : num 1 2 3 4 5 6  
## $ Spain/2003 : num 480 40 240 80 80 160  
## $ Zhejiang/07 : num 320 5 160 5 5 160  
## $ Swine/31 : logi NA NA NA NA NA NA  
## $ Illinois/09 : logi NA NA NA NA NA NA  
## $ Minnesota/09: logi NA NA NA NA NA NA  
## $ Nebraska/13 : logi NA NA NA NA NA NA  
## $ Iowa/73 : num 5 5 5 5 5 5  
## $ WI/97 : num 160 40 240 80 80 160  
## $ Colorado/09 : num 640 80 960 160 120 320  
## $ NC/34543/09 : num 1920 320 1920 640 320 320  
## $ MN/15 : num 5 5 5 5 5 5  
## $ Utah/09 : logi NA NA NA NA NA NA  
## $ NC/09 : logi NA NA NA NA NA NA  
## $ Missouri/13 : num 5 5 5 5 5 5  
## $ NC/15 : logi NA NA NA NA NA NA  
## $ Indiana/00 : num 5 5 5 5 5 5  
## $ NC/01 : num 5 5 5 5 5 5  
## $ NC/5043-1/09: num 2560 640 1280 1280 320 1280

Move the monoclonal antibody concentrations and the sera titers into the same column

p1\_mab <- p1\_mab\_full %>%  
 gather(key = "antigen", value = "concentration", 2:19)  
  
ca09\_mab <- ca09\_mab\_full %>%  
 gather(key = "antigen", value = "concentration", 2:19)  
  
p1\_sera <- p1\_sera\_full %>%  
 gather(key = "antigen", value = "titer", 2:19)  
  
ca09\_sera <- ca09\_sera\_full %>%  
 gather(key = "antigen", value = "titer", 2:19)

The sera was not ran against all the viruses due to a limited supply. So the viruses that were used for sera testing are filtered below.

ca09\_sera <- ca09\_sera %>%  
 na.omit()  
p1\_sera <- p1\_sera %>%  
 na.omit()

The antibody concentrations are in two fold dilutions, the titers will be log2(value)

p1\_mab$log2 <- log2(p1\_mab$concentration)  
p1\_mab$dilution <- log2(1/(p1\_mab$concentration/20))\*2  
  
p1\_sera$log2 <- log2(p1\_sera$titer)  
  
  
ca09\_mab$log2 <- log2(ca09\_mab$concentration)  
ca09\_mab$dilution <- log2(1/(ca09\_mab$concentration/20))\*2  
  
ca09\_sera$log2 <- log2(ca09\_sera$titer)

#The lineages of the viruses need to be added. The antigen\_key contains all the information about the viruses.

#add the antigen key to the sera data for switch out the antigen with the antigen key for ease of script writing  
  
ca09\_sera <- ca09\_sera %>%  
 left\_join(antigen\_key, by = c("antigen" = "short\_name"))  
  
p1\_sera <- p1\_sera %>%  
 left\_join(antigen\_key, by = c("antigen" = "short\_name"))  
  
ca09\_mab <- ca09\_mab %>%  
 left\_join(antibody\_key, by = "clone\_name")  
  
p1\_mab <- p1\_mab %>%  
 left\_join(antibody\_key, by = "clone\_name")

save data as RDS

saveRDS(ca09\_mab, file = "~/Documents/Data Science/AmandaSkarlupka-Project/data/processed\_data/ca09mab\_processed\_data.rds")  
saveRDS(ca09\_sera, file = "~/Documents/Data Science/AmandaSkarlupka-Project/data/processed\_data/ca09sera\_processed\_data.rds")  
saveRDS(p1\_mab, file = "~/Documents/Data Science/AmandaSkarlupka-Project/data/processed\_data/p1mab\_processed\_data.rds")  
saveRDS(p1\_sera, file = "~/Documents/Data Science/AmandaSkarlupka-Project/data/processed\_data/p1sera\_processed\_data.rds")  
#saveRDS(p1\_sera\_full, file = "./data/processed\_data/p1sera\_full\_processed\_data.rds")  
#saveRDS(CA09\_sera\_full, file = "./data/processed\_data/CA09sera\_full\_processed\_data.rds")  
#saveRDS(p1\_mabs\_full, file = "./data/processed\_data/p1mabs\_full\_processed\_data.rds")  
#saveRDS(CA09\_mabs\_full, file = "./data/processed\_data/CA09mabs\_full\_processed\_data.rds")