class 18

Jaewon Kim

Install datapasta package

```
# install.packages("datapasta")
```

Scrape Pertussis data from CDC using datapasta

```
cdc <- data.frame(
                          Year = c(1922L,
                                   1923L,1924L,1925L,1926L,1927L,1928L,
                                   1929L,1930L,1931L,1932L,1933L,1934L,1935L,
                                   1936L,1937L,1938L,1939L,1940L,1941L,
                                   1942L,1943L,1944L,1945L,1946L,1947L,1948L,
                                   1949L,1950L,1951L,1952L,1953L,1954L,
                                   1955L,1956L,1957L,1958L,1959L,1960L,
                                   1961L,1962L,1963L,1964L,1965L,1966L,1967L,
                                   1968L,1969L,1970L,1971L,1972L,1973L,
                                   1974L,1975L,1976L,1977L,1978L,1979L,1980L,
                                   1981L,1982L,1983L,1984L,1985L,1986L,
                                   1987L,1988L,1989L,1990L,1991L,1992L,1993L,
                                   1994L,1995L,1996L,1997L,1998L,1999L,
                                   2000L,2001L,2002L,2003L,2004L,2005L,
                                   2006L,2007L,2008L,2009L,2010L,2011L,2012L,
                                   2013L,2014L,2015L,2016L,2017L,2018L,
                                   2019L,2020L,2021L),
  No..Reported.Pertussis.Cases = c(107473,
                                   164191,165418,152003,202210,181411,
                                   161799,197371,166914,172559,215343,179135,
                                   265269,180518,147237,214652,227319,103188,
```

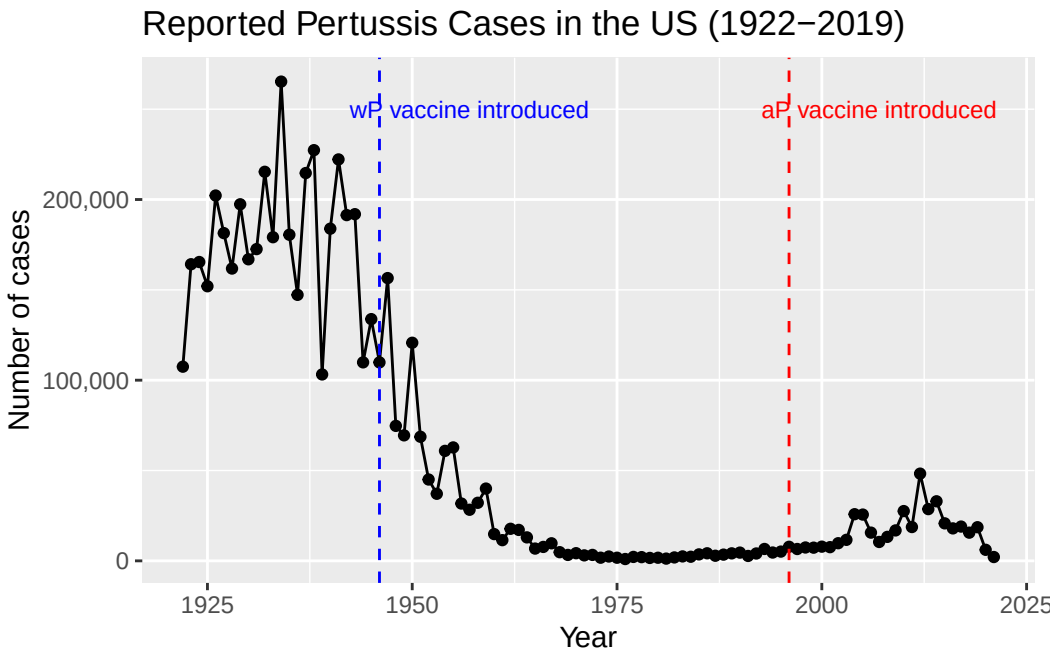
```
                                    183866,222202,191383,191890,109873,
                                    133792,109860,156517,74715,69479,120718,
                                    68687,45030,37129,60886,62786,31732,28295,
                                    32148,40005,14809,11468,17749,17135,
                                    13005,6799,7717,9718,4810,3285,4249,
                                    3036,3287,1759,2402,1738,1010,2177,2063,
                                    1623,1730,1248,1895,2463,2276,3589,
                                    4195,2823,3450,4157,4570,2719,4083,6586,
                                    4617,5137,7796,6564,7405,7298,7867,
                                    7580,9771,11647,25827,25616,15632,10454,
                                    13278,16858,27550,18719,48277,28639,
                                    32971,20762,17972,18975,15609,18617,6124,
                                    2116)
)
```

Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)

plot <- ggplot(cdc, aes(x = Year, y = No..Reported.Pertussis.Cases)) +
          geom_point() +
          geom_line() +
          labs(title = "Reported Pertussis Cases in the US (1922-2019)",
               x = "Year",
               y = "Number of cases") +
          scale_y_continuous(labels = scales::comma)
```

Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```
plot +
  geom_vline(xintercept = 1946, linetype = "dashed", color = "blue") +
  annotate("text", label = "wP vaccine introduced", x = 1957, y = 250000, color = "blue",
  geom_vline(xintercept = 1996, linetype = "dashed", color = "red") +
  annotate("text", label = "aP vaccine introduced", x = 2007, y = 250000, color = "red", s
```



- Repoted case started to decrease dramatically after the introduction of aP vaccine. However, case number slowly increased after aP vaccine is introduced.

Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? - One possible explanation is that effectiveness of aP vaccine disappears faster than wP vaccine, given that it's not life-long vaccine like hepatitis B.

Overview of vaccination history

- Read wP and aP data from CMI-PB API

```
library(jsonlite)

subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
head(subject, n = 3)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          2          wP         Female Not Hispanic or Latino White
3          3          wP         Female                 Unknown White
```

```
  year_of_birth date_of_boost      dataset
1    1986-01-01    2016-09-12 2020_dataset
2    1968-01-01    2019-01-28 2020_dataset
3    1983-01-01    2016-10-10 2020_dataset
```

Q4. How many aP and wP infancy vaccinated subjects are in the dataset? Q5. How many Male and Female subjects/patients are in the dataset? Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc…)?

```
table(subject$infancy_vac)
```

```
aP wP
60 58
```

```
table(subject$biological_sex)
```

```
Female   Male
    79     39
```

```
table(subject$race, subject$biological_sex)
```

```
                                        Female Male
  American Indian/Alaska Native              0    1
  Asian                                     21   11
  Black or African American                  2    0
  More Than One Race                         9    2
  Native Hawaiian or Other Pacific Islander  1    1
  Unknown or Not Reported                   11    4
  White                                     35   20
```

There are 60/58 aP/wP vaccinated subjects, respectively. Throughout the dataset, there were 79 females and 39 males. The breakdown of race and biological sex is described on the table above.

Side note: dealing with dates

```
#install.packages("lubridate")
library(lubridate)
```

```
Warning:    'lubridate'  R    4.3.3
```

```
         : 'lubridate'
```

```
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
today() #Check today's date
```

```
[1] "2024-03-18"
```

```
today() - ymd("2001-03-20") #Calculate the period between two dates
```

```
Time difference of 8399 days
```

```
time_length(today() - ymd("2001-03-20"), "years") #Calculate the period between two dates
```

```
[1] 22.99521
```

Q7. Using this approach determine (i) the average age of wP individuals, (ii) the average age of aP individuals; and (iii) are they significantly different?

```
library(dplyr)
```

```
         : 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
#Calculate ages and add into the data frame
subject$age <- today() - ymd(subject$year_of_birth)
subject$age_year <- time_length(subject$age, "years")

#Find statistics of age for both vaccination group
ap <- subject %>%
  filter(infancy_vac == "aP")
round(summary( time_length( ap$age, "years")))
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   21      26      26      26      27      30
```

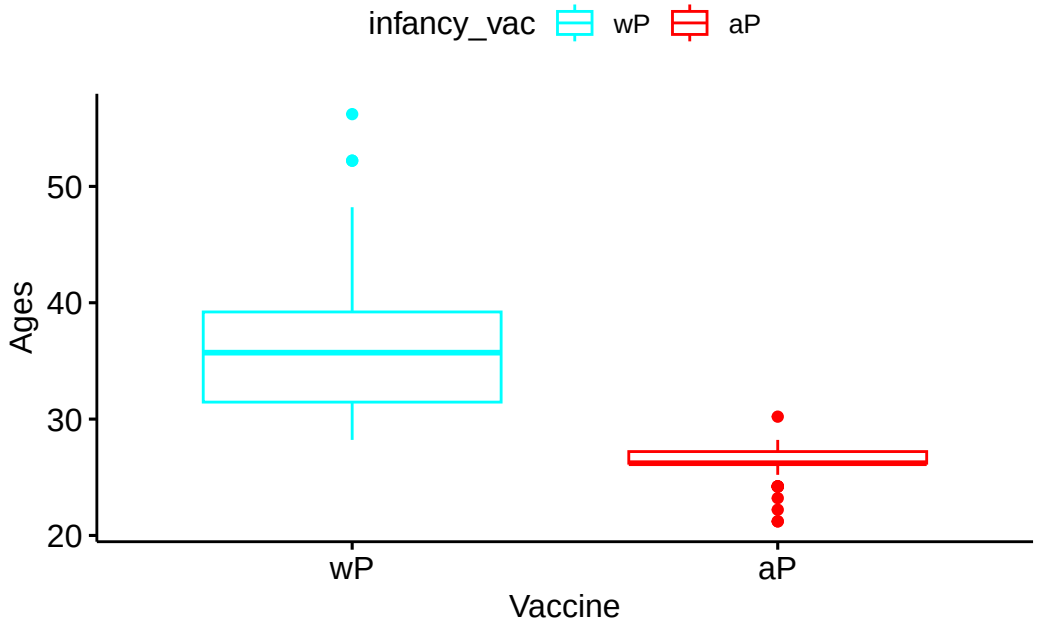
```
wp <- subject %>%
  filter(infancy_vac == "wP")
round(summary( time_length( wp$age, "years")))
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   28      31      36      37      39      56
```

```
#Plot the data and execute two-sample t-test
library("ggpubr")
```

```
Warning:    'ggpubr'  R   4.3.3
```

```
ggboxplot(subject, x = "infancy_vac", y = "age_year",
          color = "infancy_vac", palette = c("cyan", "Red"),
        ylab = "Ages", xlab = "Vaccine")
```




```
t.test(time_length(ap$age, "years"), time_length(wp$age, "years"))
```

```
    Welch Two Sample t-test

data:  time_length(ap$age, "years") and time_length(wp$age, "years")
t = -12.436, df = 65.411, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.950080  -8.643385
sample estimates:
mean of x mean of y
 26.30956  36.60629
```

Average age of aP vaccinated individuals was 26.3yrs old, where age of wP vaccinated individuals was 36.6yrs old. Since 95% CI of '11.95~-8.64 does not contain zero and p-value < 0.05, therefore, null hyphothesis is rejected. Hence, two average ages are statistically significantly different.

Q8. Determine the age of all individuals at time of boost?

```
days_at_boost <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(days_at_boost, "years")
head(age_at_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

```
summary(age_at_boost)
```

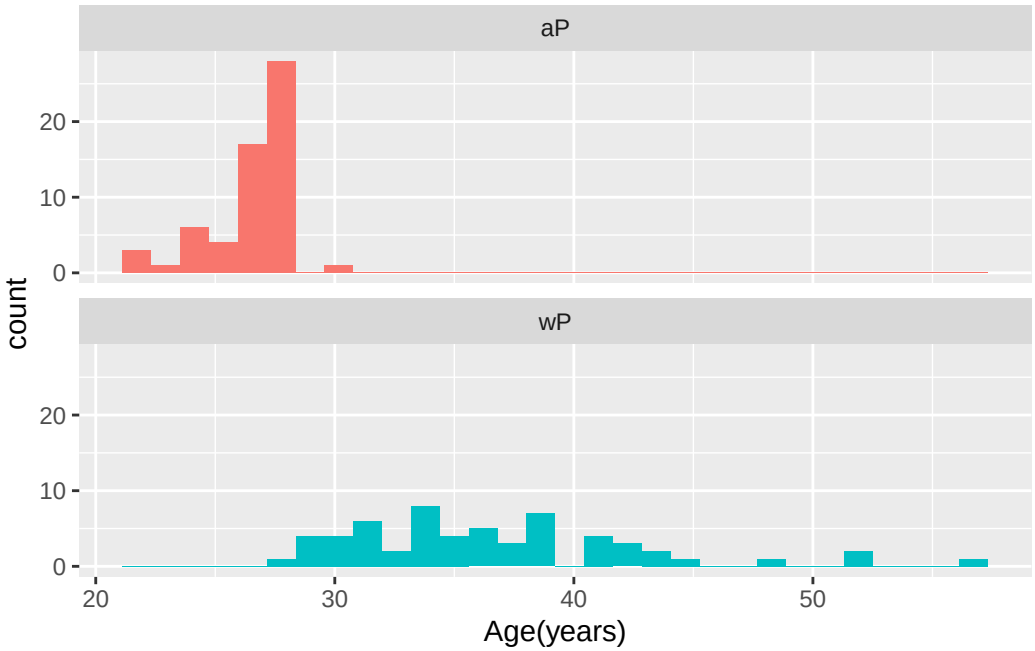
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.83   20.08   24.11   25.67   28.88   51.07
```

Ages of boost were spread among age of 18 to 51, where average was 25.67yrs old.

Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different?

```
ggplot(subject) +
  aes(time_length(age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac), nrow = 2) +
  xlab("Age(years)")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Two distribution barely overlaps, indicating that two groups are significantly different.

Joining specimen and titer data

```
#Read API data
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/plasma_ab_titer", simplifyVector = TRUE)
head(specimen, n = 3)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
  planned_day_relative_to_boost specimen_type visit
1                             0         Blood     1
2                             1         Blood     2
3                             3         Blood     3
```

```
head(titer, n = 3)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
   unit lower_limit_of_detection
1 UG/ML                2.096133
2 IU/ML               29.170000
3 IU/ML                0.530000
```

Q9. Complete the code to join specimen and subject tables to make a new merged data frame containing all specimen records along with their associated subject details:

```r
meta <- left_join(specimen, subject)
```

```
Joining with `by = join_by(subject_id)`
```

```r
dim(meta)
```

```
[1] 939  15
```

```r
head(meta)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
5           5          1                           11
6           6          1                           32
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             1         Blood     2          wP         Female
3                             3         Blood     3          wP         Female
4                             7         Blood     4          wP         Female
5                            14         Blood     5          wP         Female
6                            30         Blood     6          wP         Female
              ethnicity  race year_of_birth date_of_boost      dataset
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
```

```
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
         age age_year
1 13956 days 38.20945
2 13956 days 38.20945
3 13956 days 38.20945
4 13956 days 38.20945
5 13956 days 38.20945
6 13956 days 38.20945
```

Q10. Now using the same procedure join meta with titer data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

```
abdata <- inner_join(titer, meta, by = join_by(specimen_id))
dim(abdata)
```

```
[1] 46906    22
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
table(abdata$isotype)
```

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 4255 8983 8990 8990 8990
```

Therefore 6698, 4255, 8983, 8990, 8990, and 8990 specimens collected for IgE, IgG, IgG1, IgG2, IgG3, and IgG4, respectively.

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?

```
table(abdata$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset
       31520         8085         7301
```

$dataset indicates year of spiecemen was collected. Most recent dataset is 2022, and collected spiecemen count is the lowest among 3 years.

#Examine IgG AB titer level

```
#Filter IgG data
igg <- abdata %>% filter(isotype == "IgG")
tail(igg)
```

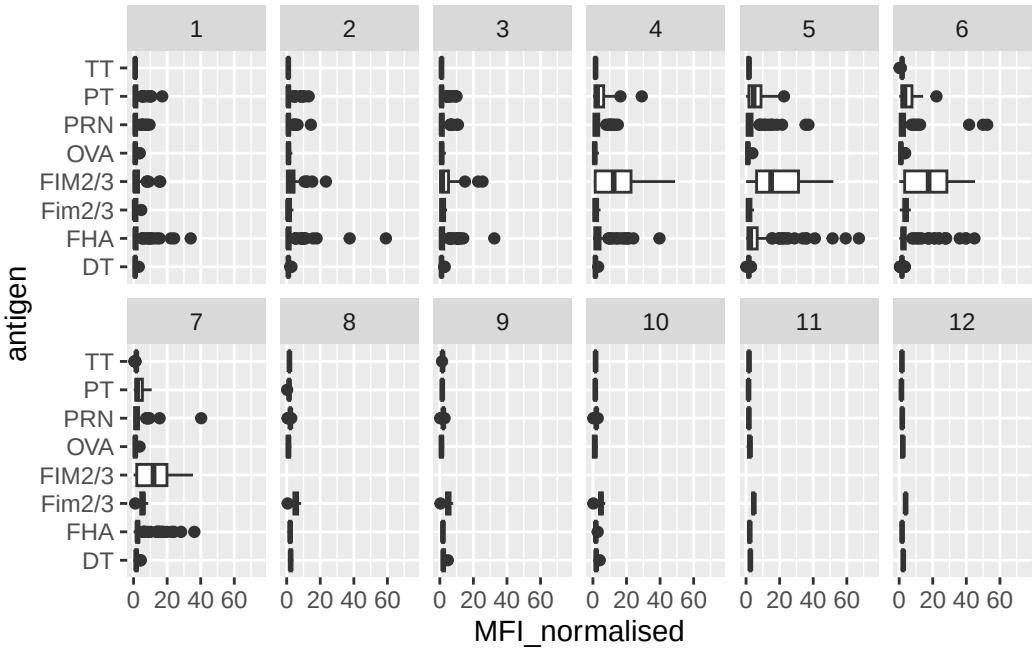
```
     specimen_id isotype is_antigen_specific antigen      MFI MFI_normalised
4250         789     IgG                TRUE      TT 25610.05       1.544882
4251         828     IgG                TRUE      TT 24288.30       1.465150
4252         858     IgG                TRUE      TT 21213.90       1.279692
4253         848     IgG                TRUE      TT 23016.15       1.388409
4254         749     IgG                TRUE      TT 23906.60       1.442124
4255         838     IgG                TRUE      TT 25658.90       1.547829
     unit lower_limit_of_detection subject_id actual_day_relative_to_boost
4250  MFI                0.4564662        102                           123
4251  MFI                0.4564662        106                           134
4252  MFI                0.4564662        109                           151
4253  MFI                0.4564662        108                           155
4254  MFI                0.4564662         98                           157
4255  MFI                0.4564662        107                           182
     planned_day_relative_to_boost specimen_type visit infancy_vac
4250                           120         Blood    10          aP
4251                           120         Blood    10          aP
4252                           120         Blood    10          wP
4253                           120         Blood    10          wP
4254                           120         Blood    12          wP
4255                           120         Blood    10          aP
     biological_sex              ethnicity  race year_of_birth date_of_boost
4250           Male Not Hispanic or Latino White    2003-01-01    2021-11-01
4251         Female Not Hispanic or Latino White    1996-01-01    2021-09-07
4252         Female Not Hispanic or Latino White    1989-01-01    2021-09-27
4253         Female Not Hispanic or Latino White    1995-01-01    2021-09-27
4254         Female Not Hispanic or Latino White    1993-01-01    2021-09-27
4255         Female Not Hispanic or Latino Asian    1998-01-01    2021-09-07
          dataset        age age_year
4250 2022_dataset  7747 days 21.21013
4251 2022_dataset 10304 days 28.21081
4252 2022_dataset 12860 days 35.20876
4253 2022_dataset 10669 days 29.21013
```

```
4254 2022_dataset 11399 days 31.20876
4255 2022_dataset  9573 days 26.20945
```

Q13. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
  xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

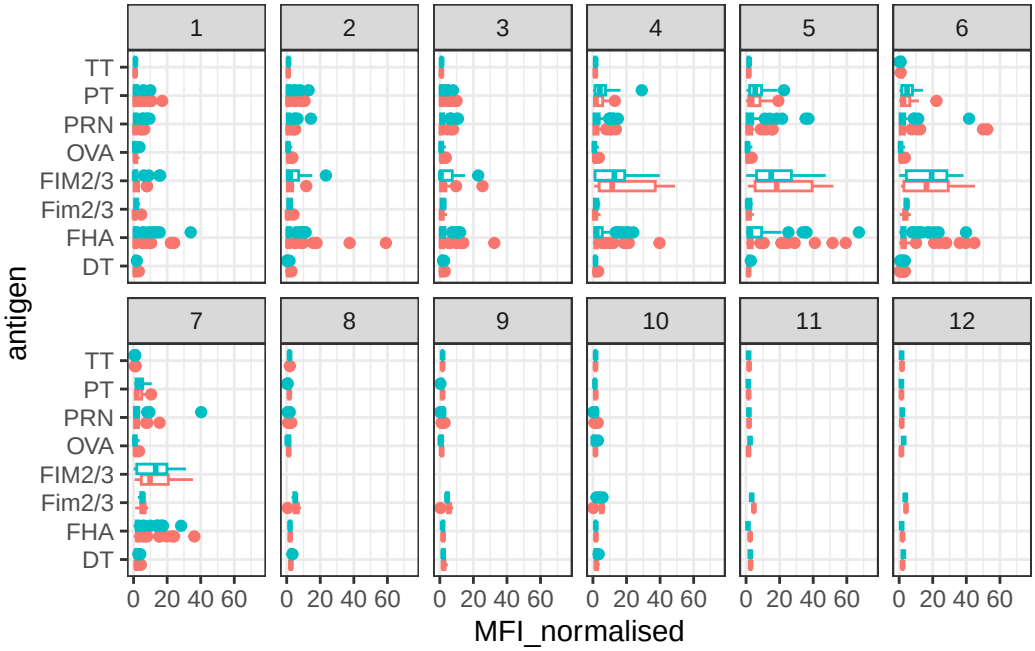
```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others? - While MFI level of other antigens stays relatively constant, level of FIM2/3 increased significantly from visit 1 to 7. This is due to affect of booster, as vaccine increase antibody responses to selected antigens. In this case, vaccine likely contains FIM2/3 antigen as main component.

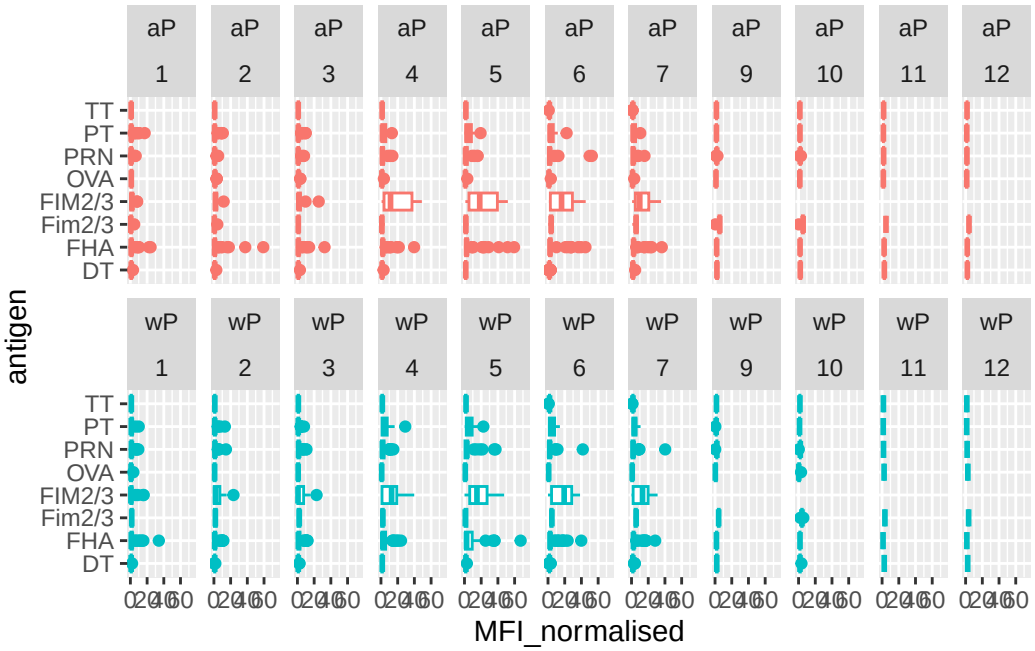
```
#Distinguish graph with types of vaccines used

#By visit
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



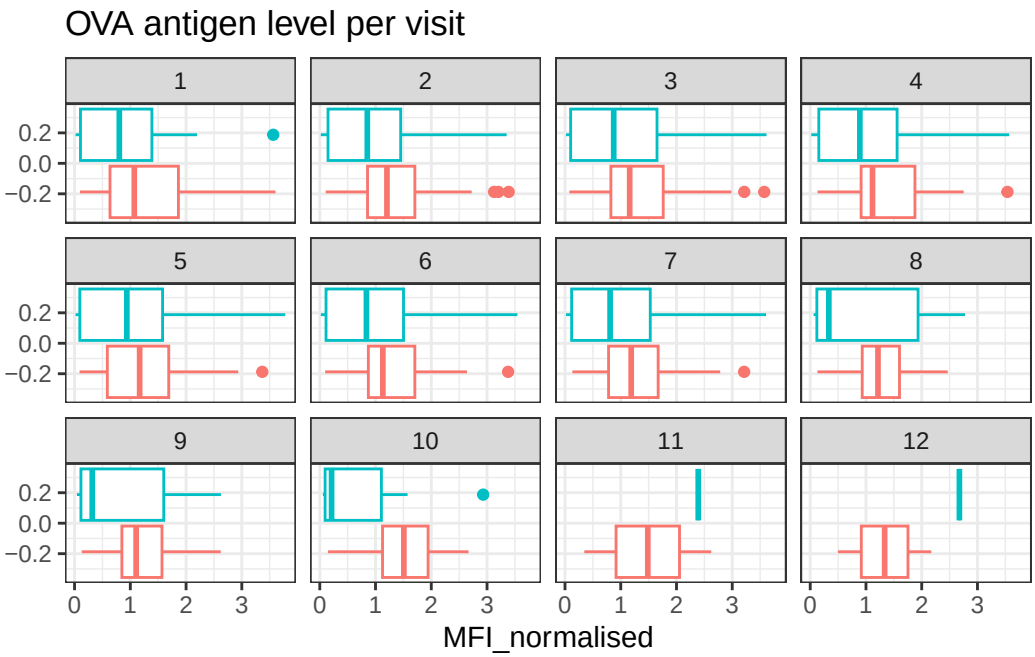

```
#By visit and vaccine type
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

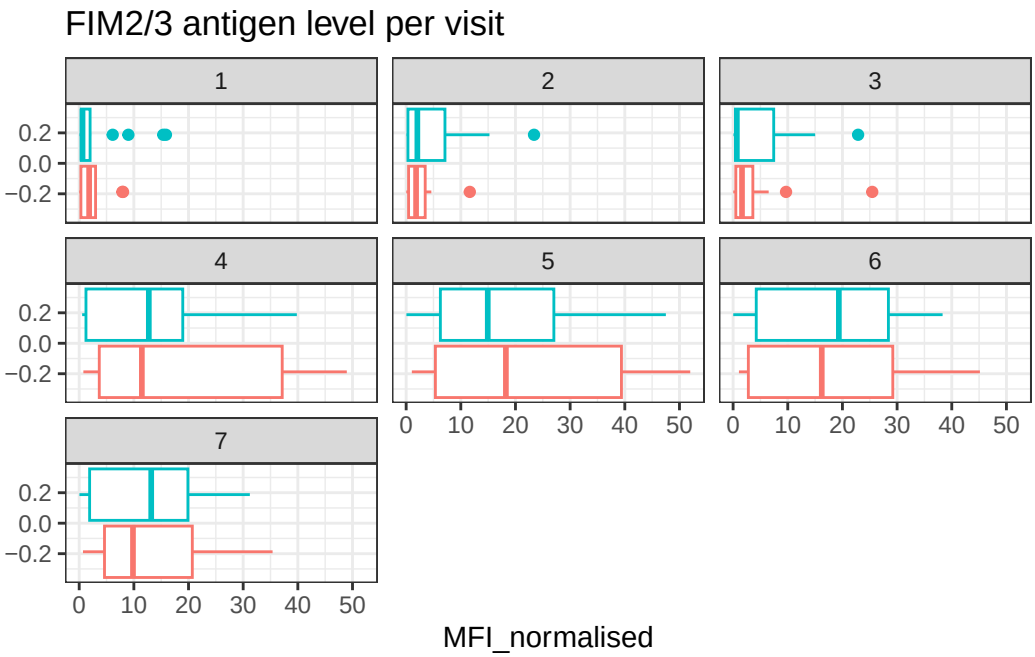


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium B. pertussis).

```
filter(igg, antigen == "OVA") %>%
  ggplot() +
  aes(MFI_normalised, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "OVA antigen level per visit")
```




```
filter(igg, antigen == "FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "FIM2/3 antigen level per visit")
```

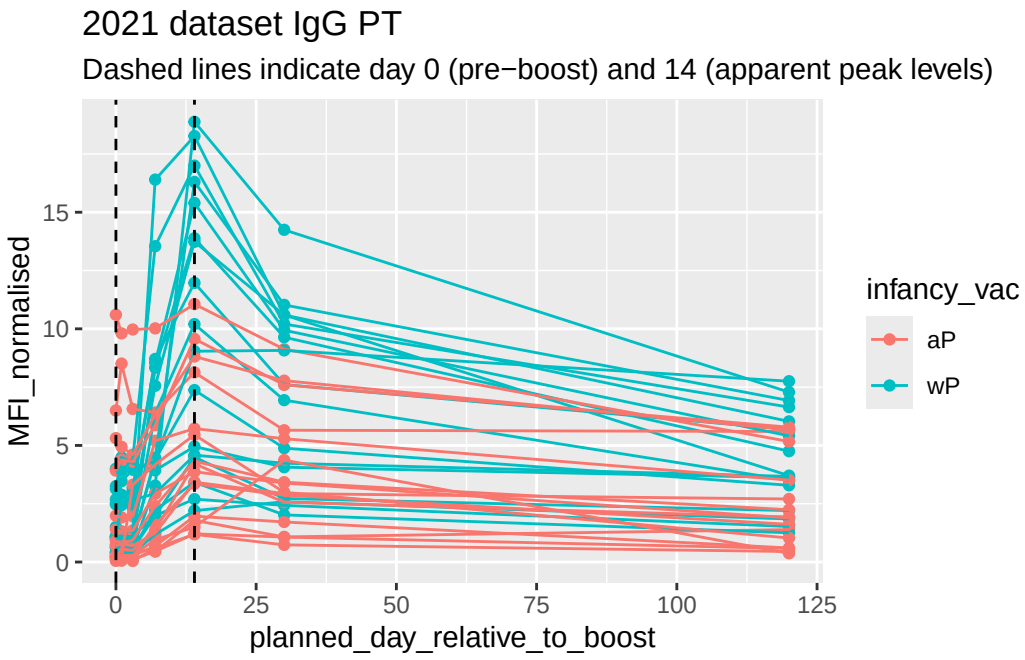


Q16. What do you notice about these two antigens time courses and the PT data in particular? - FMI level of OVA, control, stays relatively constant throughout visit. On the other hand, antigen FIM2/3 level increased significantly until visit 5 and showed mild decrease from visit 5 to 7

Q17. Do you see any clear difference in aP vs. wP responses? - While FMI level of aP vaccinated individuals showed lower response than wP from visit 6, visit 1 to 5 did not show significant difference between two groups. Thereforem there is no clear difference in vaccine response.

```
#Wrap up examining by tracking IgG level over time

abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y = MFI_normalised,
        col = infancy_vac,
        group = subject_id) +
    geom_point() +
```

```
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title = "2021 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```

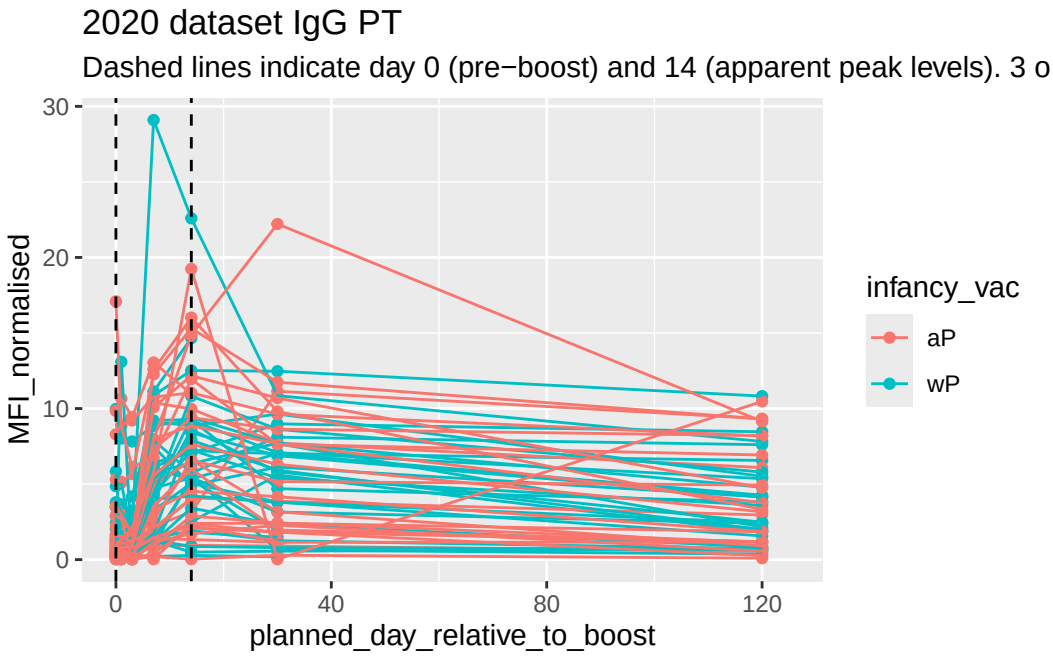



```
abdata.20 <- abdata %>% filter(dataset == "2020_dataset")

abdata.20%>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    aes(x = planned_day_relative_to_boost,
        y = MFI_normalised,
        col = infancy_vac,
        group = subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title = "2020 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels).
```

```
  xlim(0, 125)
```

```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_point()`).
```

```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_line()`).
```



2021 and 2020 datasets are similar in a sense that FMI levels spikes until day 14 (dashed line) and then gradually decrease in both. However, wP vaccine tends to show higher FMI over time in 2021 data, while aP and wP shows similar performance in 2020 data.

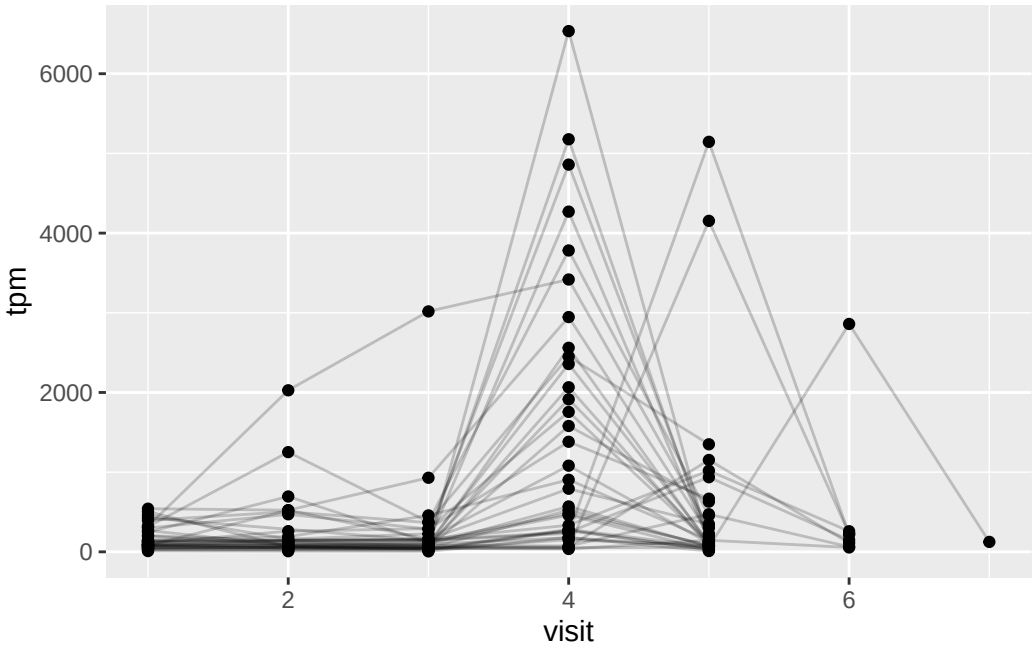
#Obtaining CMI-PB RNAseq data

```
#Read api data
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.
rna <- read_json(url, simplifyVector = TRUE)

#Join RNA data with metadata
ssrna <- inner_join(rna, meta)
```

```
Joining with `by = join_by(specimen_id)`
```

Q19. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm)

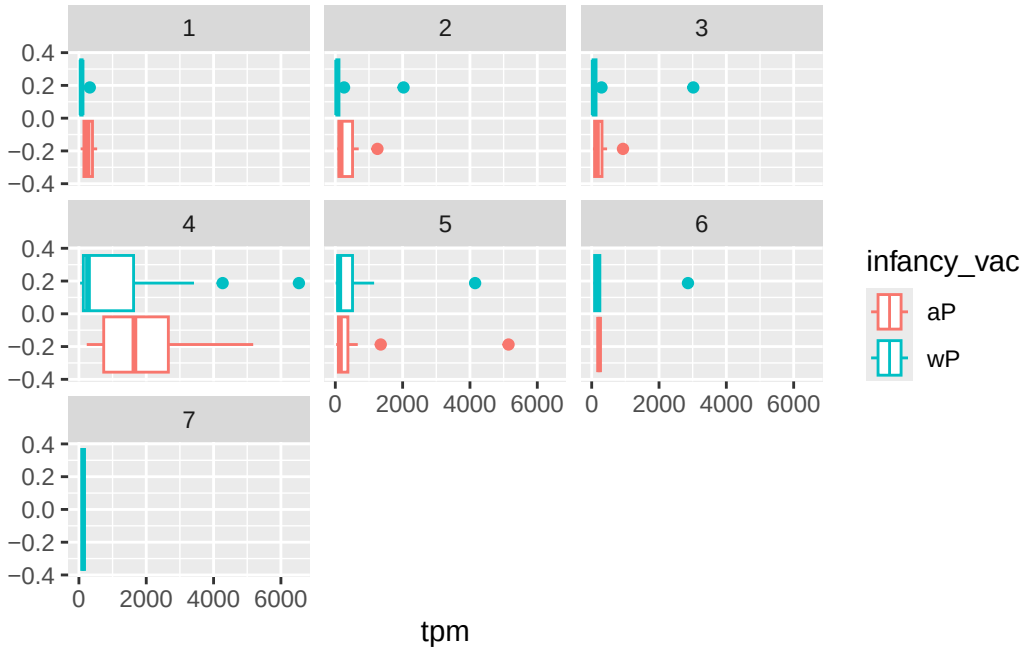
```
ggplot(ssrna) +
  aes(x = visit, y = tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



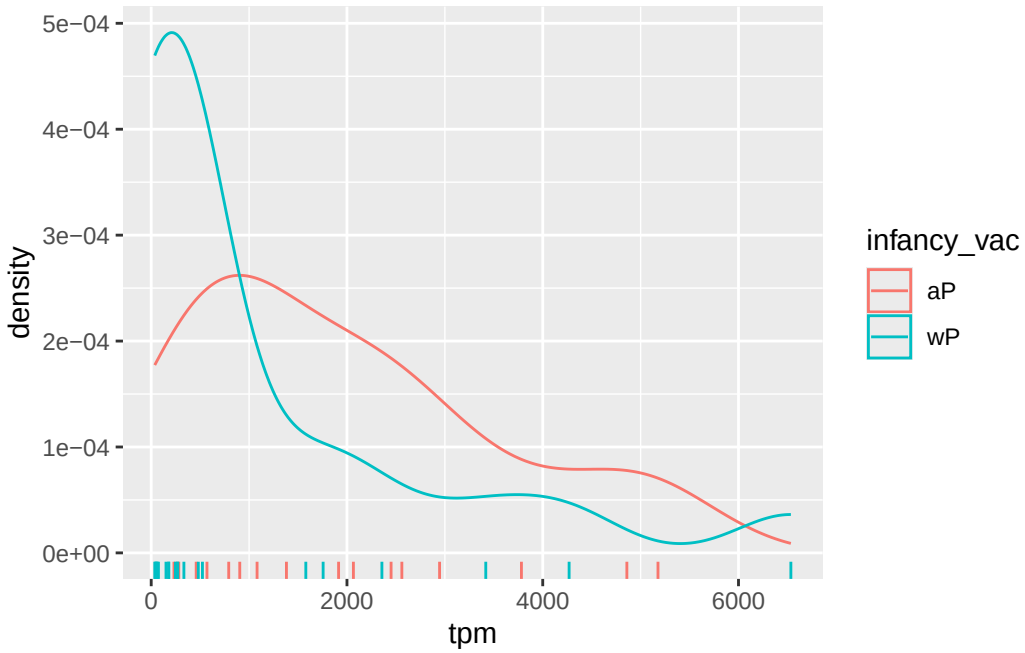
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)? - Gene expression is maximum at visit 4, when MFI level in FIM2/3 started to show significant increase. After visit 4, frequency of gene expression spike decreased.

Q21. Does this pattern in time match the trend of antibody titer data? If not, why not? - Antibody is long-lived cell, which doesn't require constant production. Therefore, single high gene expression will likely maintain high antibody response for a long time period. In graph above, gene expression spikes a single time at visit 4, which is when MFI level increases. Afterward, gene expression stays relatively low while MFI level stays high. That said, gene expression pattern matches trend of antibody titer data.

```
#Visualize tpm data. One for overall, one for visit that showed expression spike

ggplot(ssrna) +
  aes(tpm, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```




```
ssrna %>%
  filter(visit ==4 ) %>%
  ggplot() +
    aes(tpm, col = infancy_vac) + geom_density() +
    geom_rug()
```



As conclusion, aP and wP vaccine did not show significant difference in antibody expresison.