# lab19

Zijing

2022-12-02

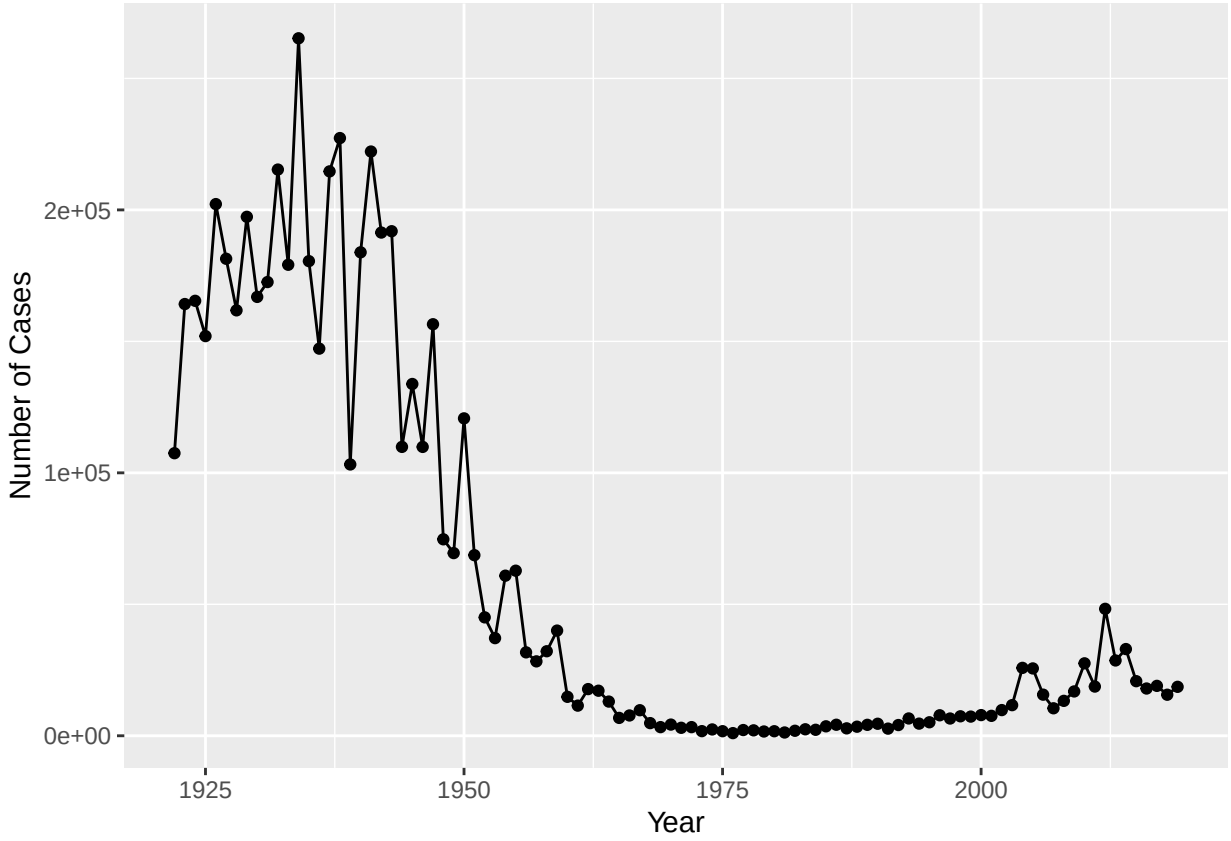
## Q1

```
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),
  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)
)
```

```
library(ggplot2)
```

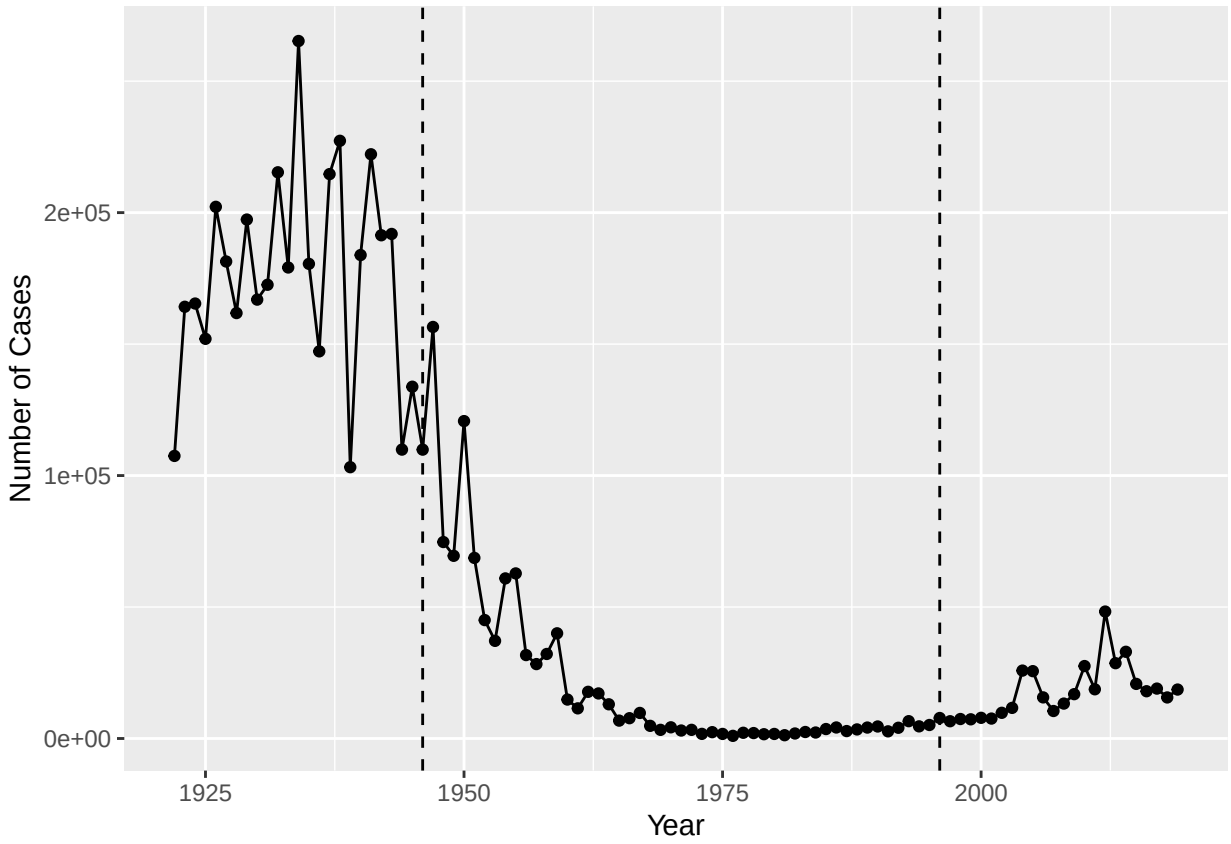
```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
```

```
  geom_point() +
  geom_line() +
  labs(x="Year",y="Number of Cases")
```




# Q2

```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept=c(1946,1996),linetype="dashed")+
  labs(x="Year",y="Number of Cases")
```




The number of cases dropped after the introduction of wP vaccination but started to rise again several years after the introduction of aP vaccination.

## Q3

The number of cases started to rise again several years after the introduction of aP vaccination. Maybe this is caused by the infection happening in grown-up infants that got aP vaccination?

```
library(jsonlite)
```

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

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

```
##
## aP wP
## 47 49
```

47 aP and 49 wP.

# Q5

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

```
##
## Female   Male
##     66     30
```

66 Female and 30 Male.

```
table(subject$race)
```

```
##
##              American Indian/Alaska Native
##                                          1
##                                      Asian
##                                         27
##                  Black or African American
##                                          2
##                           More Than One Race
##                                         10
## Native Hawaiian or Other Pacific Islander
##                                          2
##                    Unknown or Not Reported
##                                         14
##                                      White
##                                         40
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

# Q6

```
male <- subject%>%filter(biological_sex == "Male")
female <- subject%>%filter(biological_sex == "Female")
```

Male race breakdown:

```
table(male$race)
```

```
##
##             American Indian/Alaska Native
##                                         1
##                                     Asian
##                                         9
##                          More Than One Race
##                                         2
## Native Hawaiian or Other Pacific Islander
##                                         1
##                    Unknown or Not Reported
##                                         4
##                                     White
##                                        13
```

Female race breakdown:

```
table(female$race)
```

```
##
##                                     Asian
##                                        18
##                  Black or African American
##                                         2
##                          More Than One Race
##                                         8
## Native Hawaiian or Other Pacific Islander
##                                         1
##                    Unknown or Not Reported
##                                        10
##                                     White
##                                        27
```

```
library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
today()
```

```
## [1] "2022-12-03"
```

```
today() - ymd("2000-01-01")
```

```
## Time difference of 8372 days
```

```
time_length( today() - ymd("2000-01-01"),  "years")
```

```
## [1] 22.92129
```

# Q7

```
subject$age <- today() - ymd(subject$year_of_birth)
ap <- subject%>%filter(infancy_vac == "aP")
wp <- subject%>%filter(infancy_vac == "wP")
time_length(mean(ap$age),"years")
```

```
## [1] 25.23908
```

```
time_length(mean(wp$age),"years")
```

```
## [1] 36.08353
```

wP average: 36 years, aP average: 25 years.The average age of wP receivers is much higher than that of aP receivers

# Q8

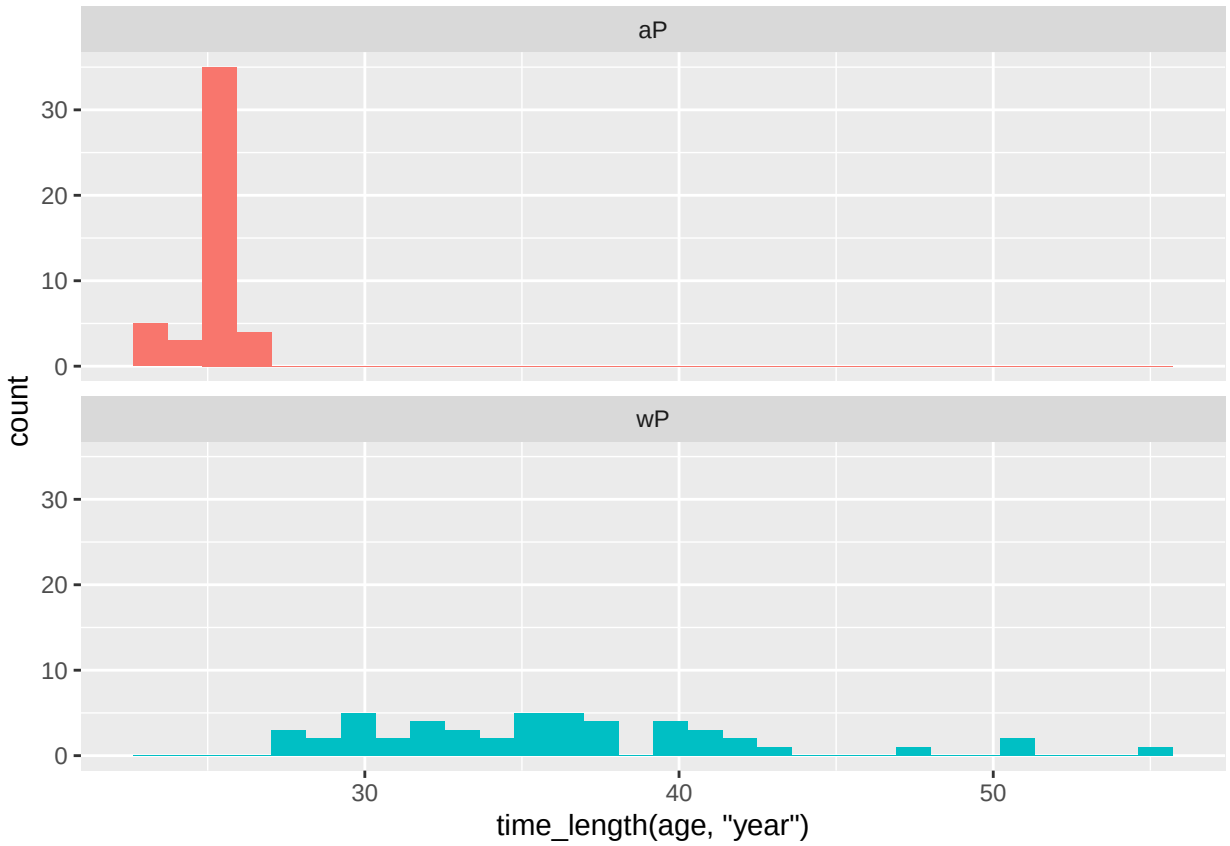
```
boost_age <- time_length(ymd(subject$date_of_boost) - ymd(subject$year_of_birth), "year")
head(boost_age)
```

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

# Q9

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

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




Yes, these two groups are different.

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
```

# Q9

```
meta <- full_join(specimen, subject)
```

```
## Joining, by = "subject_id"
```

```
dim(meta)
```

```
## [1] 729  14
```

```
head(meta)
```

```
##   specimen_id subject_id actual_day_relative_to_boost
## 1           1          1                           -3
## 2           2          1                          736
## 3           3          1                            1
## 4           4          1                            3
## 5           5          1                            7
## 6           6          1                           11
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                           736         Blood    10          wP         Female
## 3                             1         Blood     2          wP         Female
```

```
## 4                            3         Blood     3          wP         Female
## 5                            7         Blood     4          wP         Female
## 6                           14         Blood     5          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
## 1 13485 days
## 2 13485 days
## 3 13485 days
## 4 13485 days
## 5 13485 days
## 6 13485 days
```

# Q10

```
abdata <- inner_join(titer, meta)
```

```
## Joining, by = "specimen_id"
```

```
dim(abdata)
```

```
## [1] 32675    21
```

```
head(abdata)
```

```
##   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
## 4           1     IgG                TRUE     PRN  332.12718       2.602350
## 5           1     IgG                TRUE     FHA 1887.12263      34.050956
## 6           1     IgE                TRUE     ACT    0.10000       1.000000
##    unit lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1 UG/ML                2.096133          1                           -3
## 2 IU/ML               29.170000          1                           -3
## 3 IU/ML                0.530000          1                           -3
## 4 IU/ML                6.205949          1                           -3
## 5 IU/ML                4.679535          1                           -3
## 6 IU/ML                2.816431          1                           -3
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                             0         Blood     1          wP         Female
## 3                             0         Blood     1          wP         Female
## 4                             0         Blood     1          wP         Female
## 5                             0         Blood     1          wP         Female
## 6                             0         Blood     1          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
## 1 13485 days
## 2 13485 days
## 3 13485 days
## 4 13485 days
## 5 13485 days
## 6 13485 days
```

# Q11

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

```
##
##  IgE  IgG IgG1 IgG2 IgG3 IgG4
## 6698 1413 6141 6141 6141 6141
```

# Q12

```
table(abdata$visit)
```

```
##
##    1    2    3    4    5    6    7    8
## 5795 4640 4640 4640 4640 4320 3920   80
```

Much less specimens in visit 8 compared to other visits.

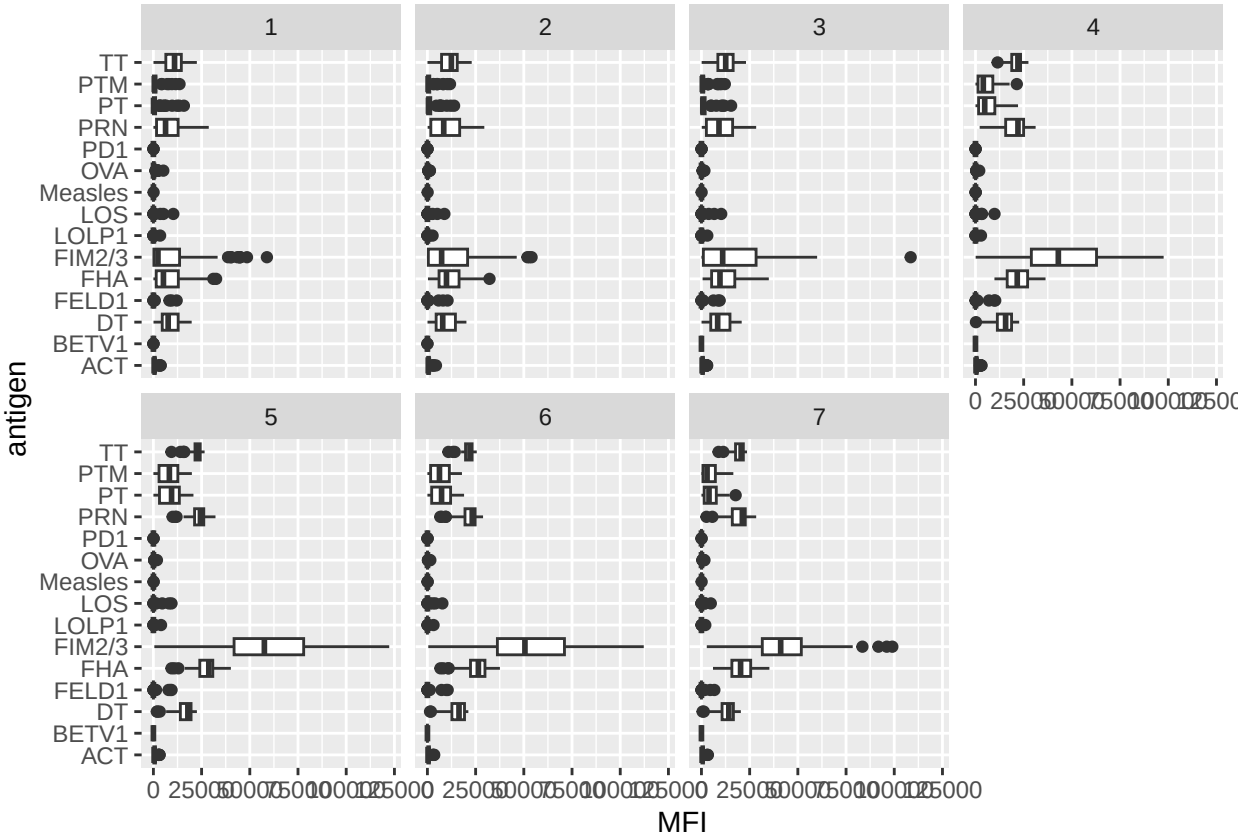
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
## 1           1    IgG1                TRUE     ACT 274.355068      0.6928058
## 2           1    IgG1                TRUE     LOS  10.974026      2.1645083
## 3           1    IgG1                TRUE   FELD1   1.448796      0.8080941
## 4           1    IgG1                TRUE   BETV1   0.100000      1.0000000
## 5           1    IgG1                TRUE   LOLP1   0.100000      1.0000000
## 6           1    IgG1                TRUE Measles  36.277417      1.6638332
##    unit lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1 IU/ML                3.848750          1                           -3
## 2 IU/ML                4.357917          1                           -3
## 3 IU/ML                2.699944          1                           -3
## 4 IU/ML                1.734784          1                           -3
## 5 IU/ML                2.550606          1                           -3
## 6 IU/ML                4.438966          1                           -3
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                             0         Blood     1          wP         Female
## 3                             0         Blood     1          wP         Female
## 4                             0         Blood     1          wP         Female
```

```
## 5                              0         Blood     1          wP         Female
## 6                              0         Blood     1          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
## 1 13485 days
## 2 13485 days
## 3 13485 days
## 4 13485 days
## 5 13485 days
## 6 13485 days
```

## Q13

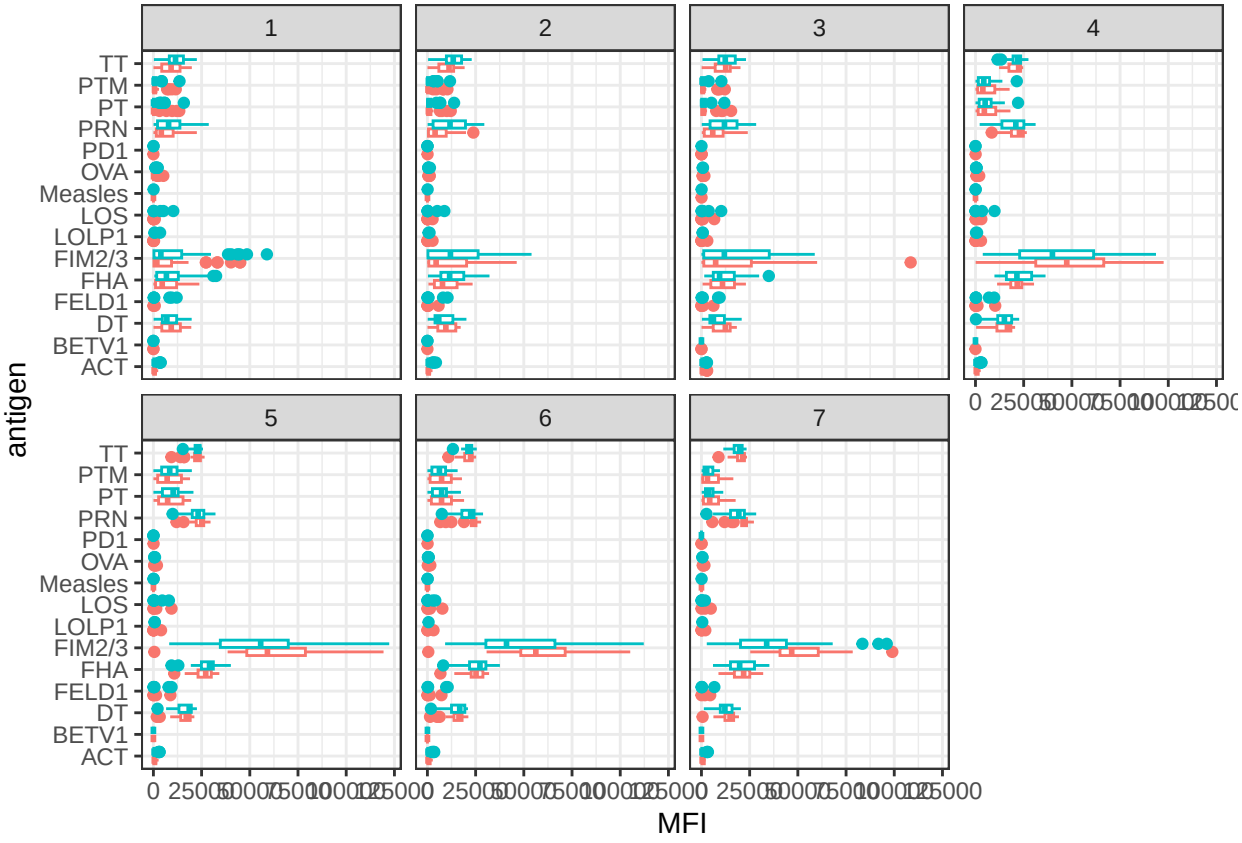
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



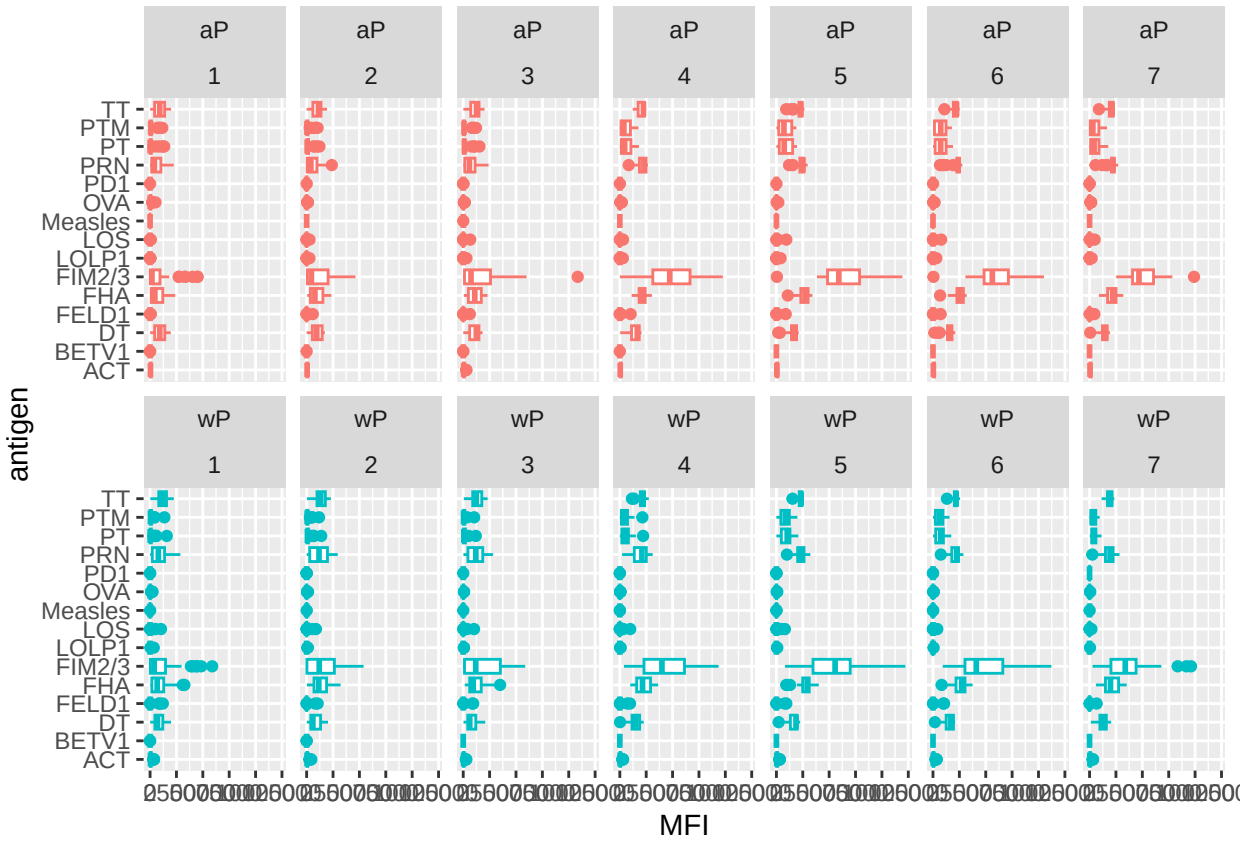

# Q14

FIM2/3 is the most different one antigen. PRN and FHA are quite different as well.

```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```

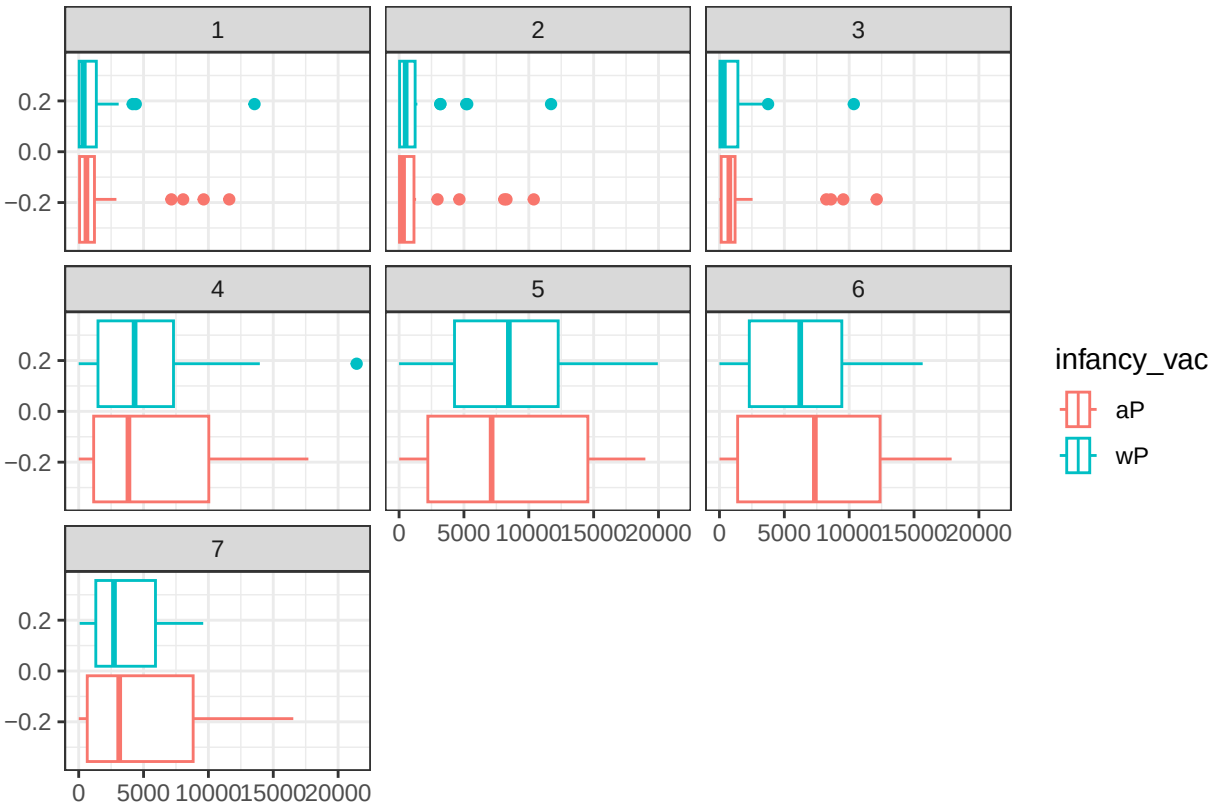



```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

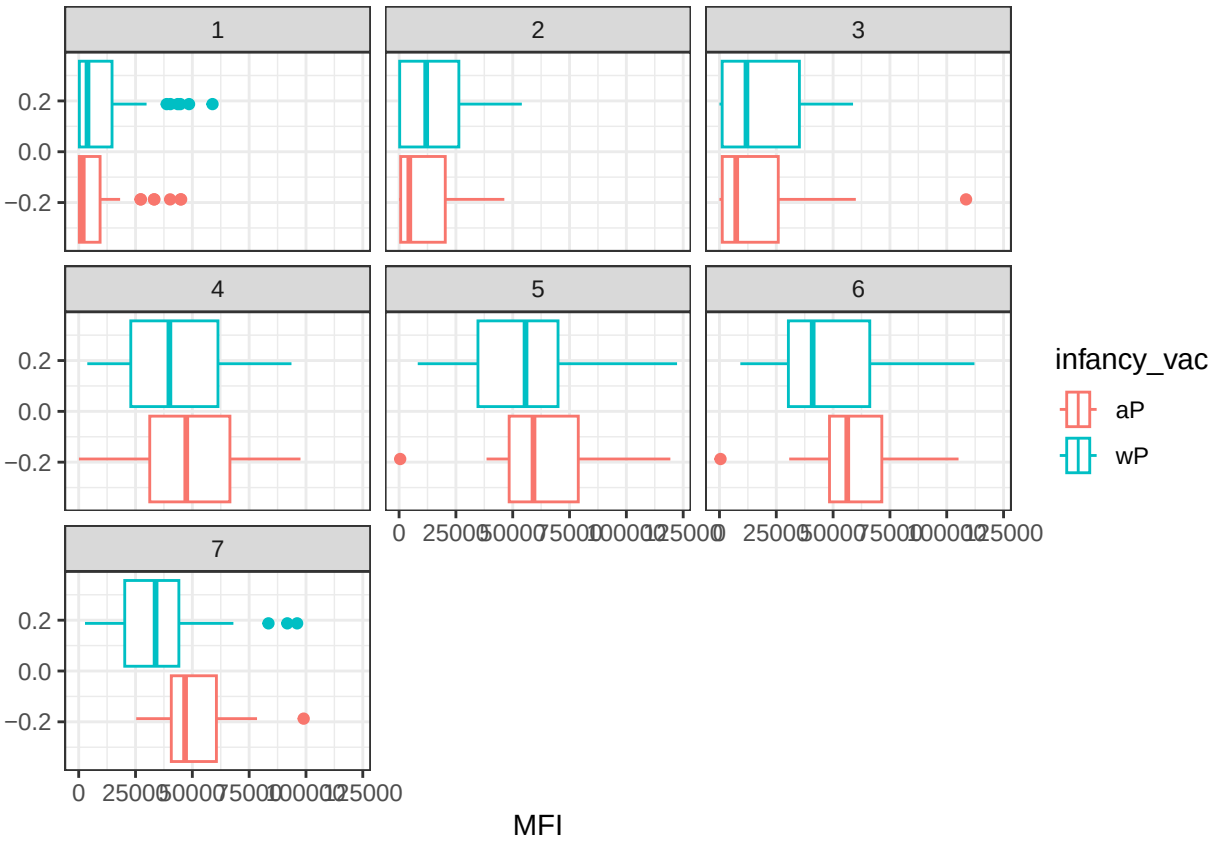



# Q15

```
filter(ig1, antigen=="PTM") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




#

Q16

Both rise over time, but FIM2/3 more significantly. Visit 5 seems to be the peak for both.

# Q17

aP response starts being lower but ends up being higher than wP response. However, they follow a similar trend of rising and then declining over time.

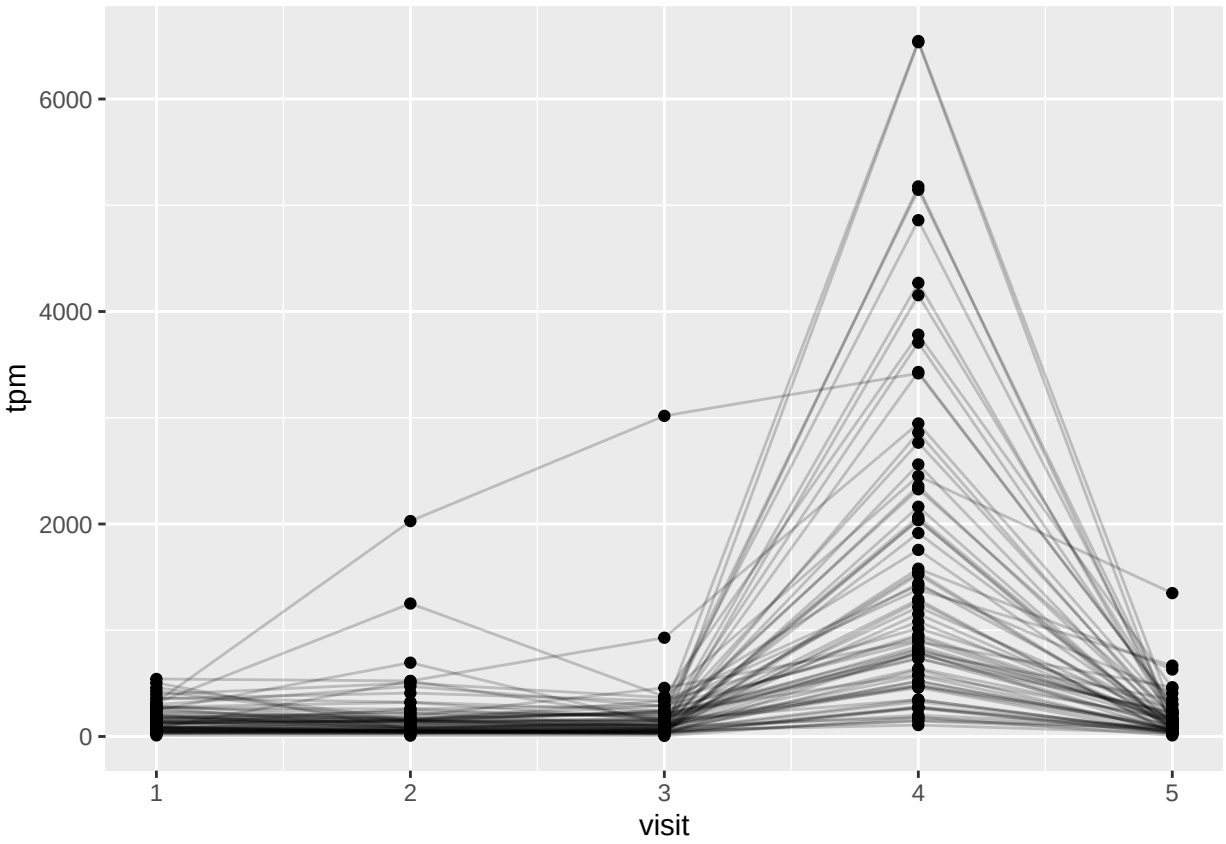
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

rna <- read_json(url, simplifyVector = TRUE)
```

```
ssrna <- inner_join(rna, meta)
```

```
## Joining, by = "specimen_id"
```

# Q18

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



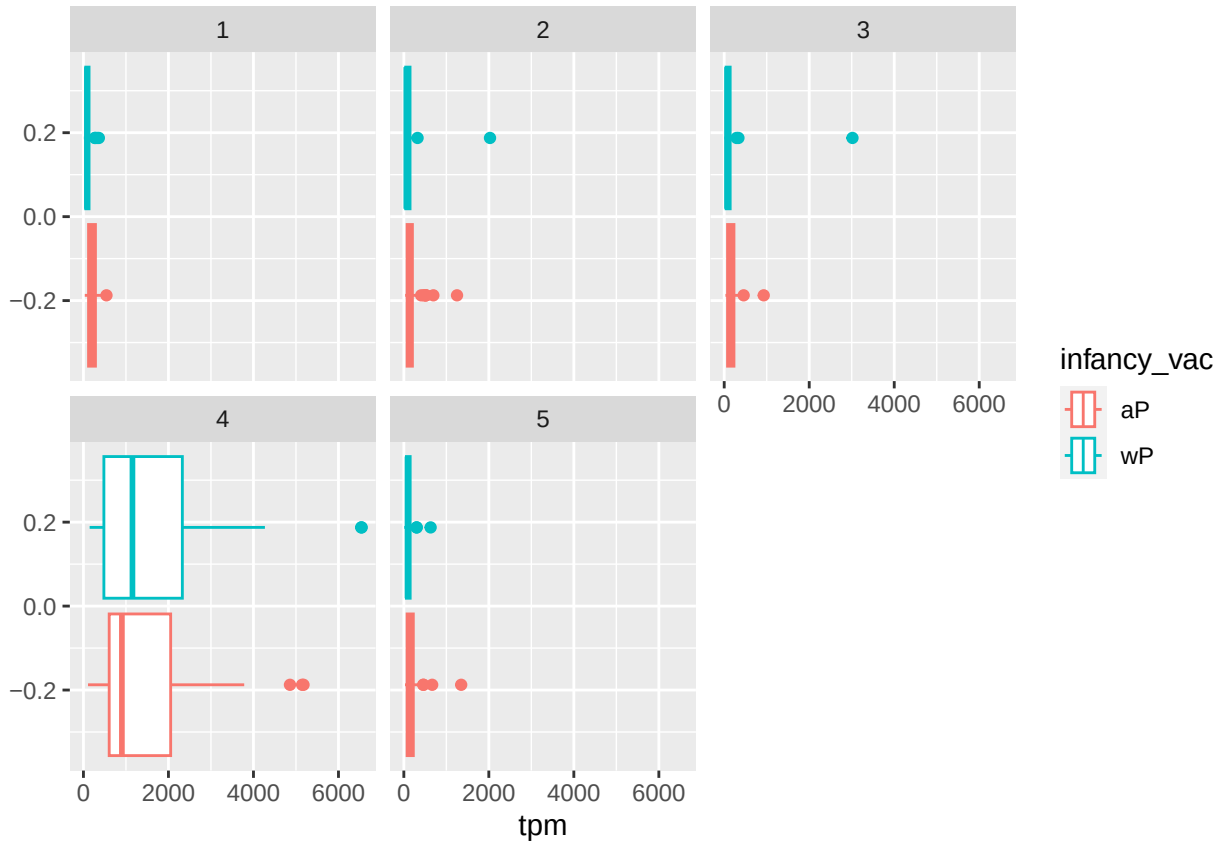

# Q19

The gene expression is at its maximum level at visit 4.

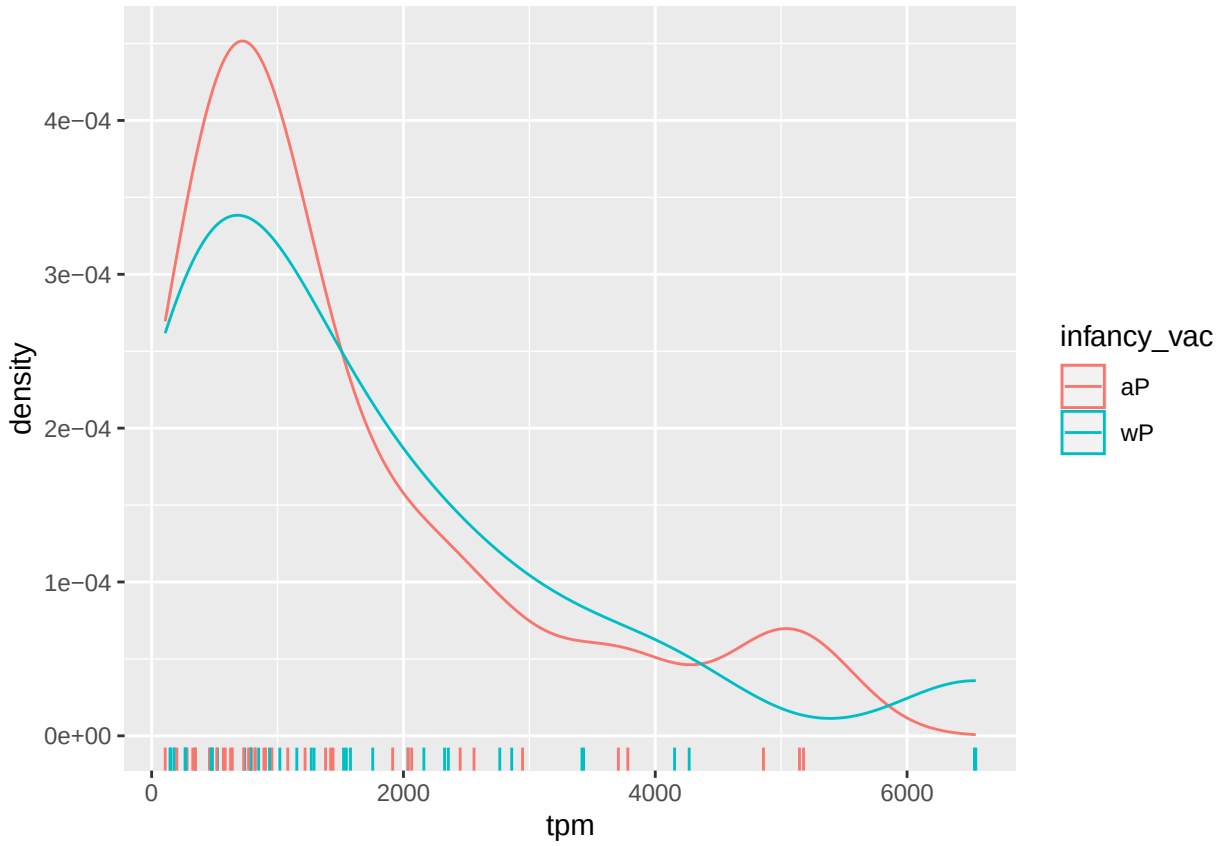
## Q20

This does not match the antibody pattern as the antibody peaks at visit 5. This is likely because antibodies are made after the genes are expressed and would live for a long time. Thus, the antibody expression continues to accumulate until gene expression has dropped to zero, which is until sometime between visit 4 and 5, leading to the peak in antibody detection on visit 5.

```
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()
```


density
4e−04
3e−04
2e−04
1e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP