Class 18 - Pertussis Mini Project

Joshua Mac

2025-03-09

Pertussis and the CMI-PB project

SECTION 1: INVESTIGATING PERTUSSIS CASES BY YEAR

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.

A1. See below.

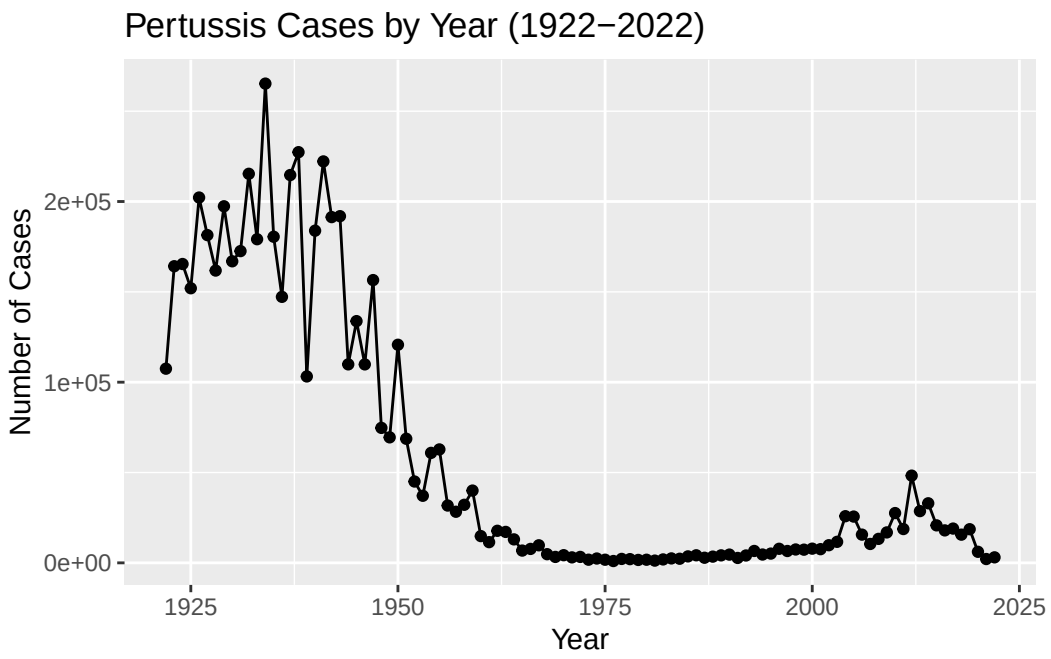
```
# Install addin "datapasta" package
# install.packages("datapasta") in console

# Assign CDC pertussis case number data to df using datapasta from addins menu
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,2022L),
        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,
                                        6124,2116,3044)
       )
```

Next, plot:

```
library(ggplot2)
cdc_plot<-ggplot(cdc, aes(Year, No..Reported.Pertussis.Cases))+
  geom_point()+
  geom_line()+
  labs(title = "Pertussis Cases by Year (1922-2022)", x = "Year", y = "Number of Cases")
cdc_plot
```

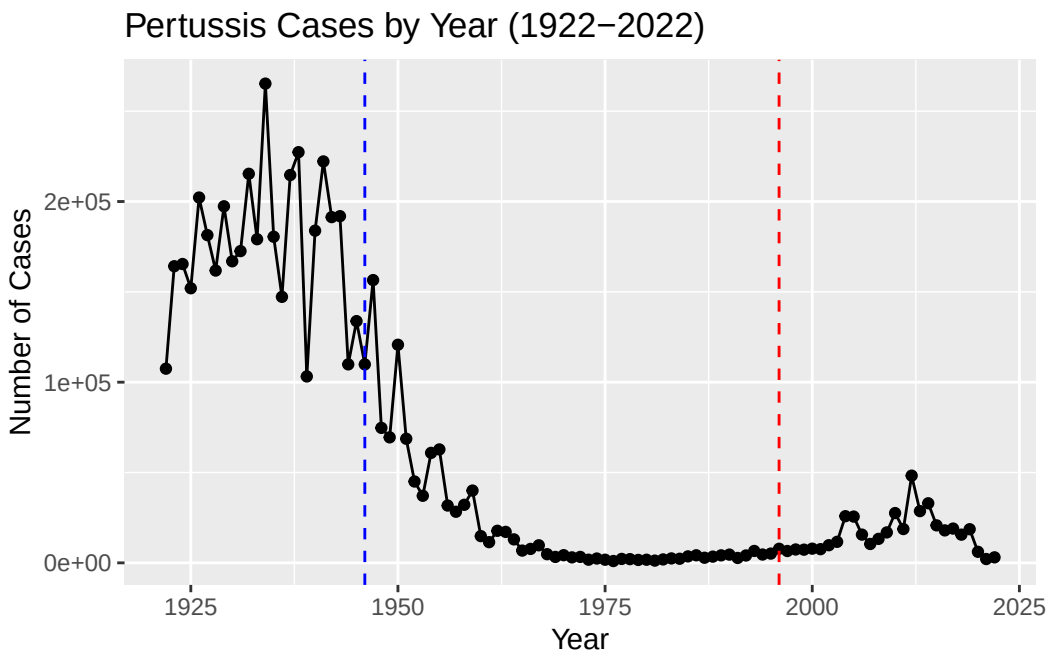


SECTION 2. A TALE OF TWO VACCINES (wP & aP)

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?

> A2. See below. I notice that with the introduction of the wP vaccine (blue line) in 1946, the number of cases of pertussis drops off significantly due to enhance immunization and preventative protection against the affliction. However, after the 1996 switch to the aP vaccine (red line), we actually see more reported cases of pertussis in the population.

```
cdc_plot +
  geom_vline(xintercept=1946, color = "blue", linetype="dashed") +
  geom_vline(xintercept=1996, color = "red", linetype="dashed")
```



Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

> A3. After the introduction of the aP vaccine, we see cases of pertussis reported rising. Most notably, in 2012, 48,277 cases are reported in the US, which has not been at such levels since 1955 (62786 cases). Clearly a public health concern, we can attribute the trend to possible mutation/evolution of the pertussis bacterium *Bordetella pertussis* and the rising resistance to vaccinations due to antivax communities and spread of misinformation…

SECTION 3. EXPLORING CMI-PB DATA

CMI-PB at a glance: Project that aims to provide to the public long-term humoral and cellular immune response data for those who received either vaccine in infancy with Tdap boosters. It aims to investigate the mechanisms underlying waning protection against pertussis by evaluating the pertussis-specific immune responses overtime as mentioned above.

The CMI-PB API returns JSON data (key-value pairs).

So, we'll need the jsonlite package.

```
# install.packages("jsonlite") in console
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. How many aP and wP infancy vaccinated subjects are in the dataset?

A4. There are 87 aP infancy vaccinated subjects and 85 wP infancy vaccinated subjects in the dataset. See below.

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

```
aP wP
87 85
```

Q5. How many Male and Female subjects/patients are in the dataset?

A5. There are 60 male and 112 female subjects in the dataset. See below.

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

```
Female   Male
   112     60
```

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

A6. There is one American Indian/Alaska Native male and no females of that race in the dataset. There are 12 Asian males and 32 Asian females in the dataset. There are three Black or African American males and two Black of African American females in the dataset. There are four males that are of More Than One Race and 15 females that are of More Than One Race in the dataset. There is one male and one female that are both Native Hawaiian or Other Pacific Islander in the dataset. There are seven males and 14 females with races unknown/not reported in the dataset. Finally, there are 32 males and 48 females that consider themselves White in the dataset.

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

```
                                        Female Male
  American Indian/Alaska Native              0    1
  Asian                                     32   12
  Black or African American                  2    3
  More Than One Race                        15    4
  Native Hawaiian or Other Pacific Islander  1    1
  Unknown or Not Reported                   14    7
  White                                     48   32
```

Side-Note for Q7 (the "lubridate" package)

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

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today() # tells today's date in "YEAR-MM-DD"
```

```
[1] "2025-03-10"
```

```
diff<- today() - ymd("2000-01-01") # matching the `today()` format, do this to find how many
diff
```

```
Time difference of 9200 days
```

```
time_length(diff, "years") # the time difference in years
```

```
[1] 25.18823
```

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?

> A7. (i) The average age of wP individuals is 36 years, (ii) aP individuals is 27, and (iii) they are significantly different as the p-value of the t-test conducted below is low (<2.2e-16) and less than alpha 0.05, suggesting statistical significance.

```
subject$age <- today() - ymd(subject$year_of_birth)
```

```
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
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      26      27      27      28      34
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     22      32      34      36      39      57 
```

```r
t.test(ap$age, wp$age, var.equal = FALSE)
```

```

    Welch Two Sample t-test

data:  ap$age and wp$age
t = -12.918 days, df = 104.03, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3686.855 days -2705.535 days
sample estimates:
Time differences in days
mean of x mean of y 
 9892.276 13088.471 
```

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

A8. See the ages, in years, at time of boost below.

```r
days_before_boost<- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(days_before_boost, "year")
head(age_at_boost, 172)
```

```
  [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
  [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
 [17] 36.69815 19.65777 22.73511 35.65777 33.65914 31.65777 25.73580 24.70089
 [25] 28.70089 33.73580 19.73443 34.73511 19.73443 28.73648 27.73443 19.81109
 [33] 26.77344 33.81246 25.77413 19.81109 18.85010 19.81109 31.81109 22.81177
 [41] 31.84942 19.84942 18.85010 18.85010 19.90691 18.85010 20.90897 19.04449
 [49] 20.04381 19.90691 19.90691 19.00616 19.00616 20.04381 20.04381 20.07940
 [57] 21.08145 20.07940 20.07940 20.07940 32.26557 25.90007 23.90144 25.90007
 [65] 28.91992 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058
 [73] 24.15058 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876
 [81] 26.20671 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375
```

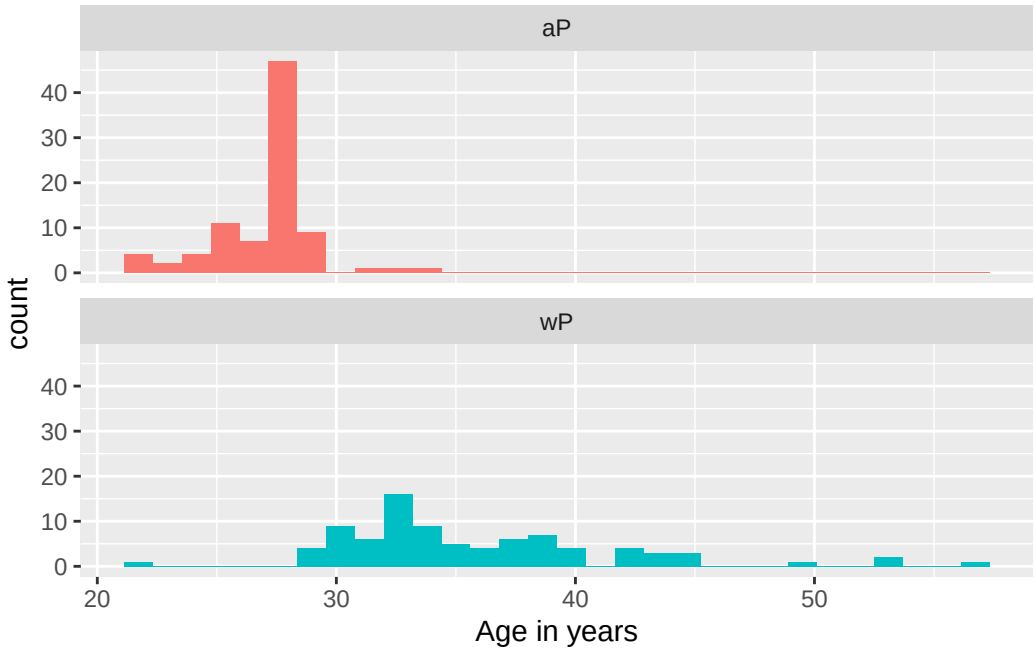
```
 [89] 22.41752 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707
 [97] 35.90965 28.73648 22.68309 20.83231 18.83368 18.83368 27.68241 32.68172
[105] 27.68241 25.68378 23.68241 26.73785 32.73648 24.73648 25.79603 25.79603
[113] 25.79603 31.79466 19.83299 21.91102 27.90965 24.06297 23.90965 27.12115
[121] 22.12183 23.12115 26.17933 22.17933 29.17728 29.23477 26.23682 28.29295
[129] 31.29363 26.29432 24.35044 27.35113 25.40999 32.41068 27.56194 27.41136
[137] 24.50650 22.56263 29.56057 21.69473 26.69678 31.90691 19.90691 23.90691
[145] 20.90623 31.00616 23.00616 35.00616 32.00548 32.00548 31.04449 28.12047
[153] 25.11978 26.11910 26.19302 22.19302 26.19302 23.19507 29.19370 27.32923
[161] 30.32717 24.55852 30.55715 32.55852 30.55715 22.67488 26.67488 32.67625
[169] 20.67625 31.75086 20.86516 36.06297
```

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

A9. We can see visually comparing histograms of the wP and aP datasets that they are different. With the t-test conducted in Q7, we are sure that statisitically they 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 in years")
```

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



Joining multiple tables

```
# Complete the API URLs...
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = TRUE)
```

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:

A9. See code below.

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

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

```
dim(meta)
```

```
[1] 1503   14
```

```
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
1 14313 days
2 14313 days
3 14313 days
4 14313 days
5 14313 days
6 14313 days
```

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.

A10. See code below.

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

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

```
dim(abdata)
```

```
[1] 61956    21
```

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

A11. For IgE there are 6698 specimens, IgG 7265 specimens, IgG1 11993 specimens, IgG2 12000 specimens, IgG3 12000 specimens, and IgG4 12000 specimens.

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

```
  IgE   IgG  IgG1  IgG2  IgG3  IgG4
 6698  7265 11993 12000 12000 12000
```

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

A12. The different $dataset values in abdata are the number of entries/reports in a given year. The number of rows for the most recent 2023 dataset doubled from the year prior, indicating a 2x jump in cases since the prior year.

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

```
2020_dataset 2021_dataset 2022_dataset 2023_dataset
       31520         8085         7301        15050
```

SECTION 4. EXAMINE IgG Ab TITER LEVELS

```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
4          19     IgG                TRUE      PT   20.11607       1.096366
5          19     IgG                TRUE     PRN  976.67419       7.652635
```

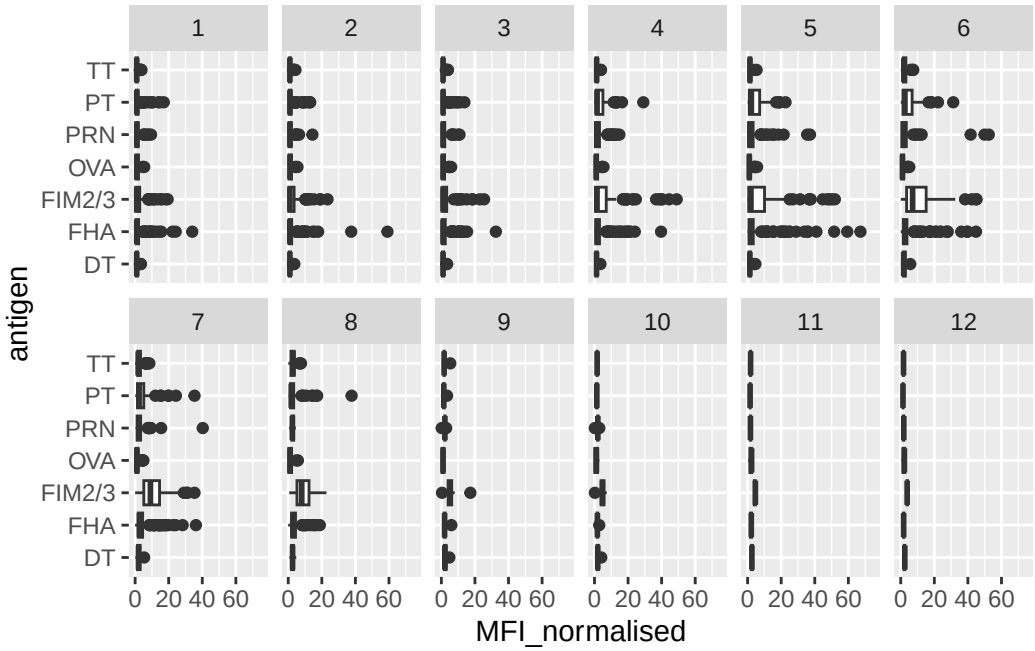
```
6          19     IgG                TRUE     FHA   60.76626       1.096457
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 0.530000          1                           -3
2 IU/ML                 6.205949          1                           -3
3 IU/ML                 4.679535          1                           -3
4 IU/ML                 0.530000          3                           -3
5 IU/ML                 6.205949          3                           -3
6 IU/ML                 4.679535          3                           -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                Unknown White    1983-01-01    2016-10-10 2020_dataset
5                Unknown White    1983-01-01    2016-10-10 2020_dataset
6                Unknown White    1983-01-01    2016-10-10 2020_dataset
         age
1 14313 days
2 14313 days
3 14313 days
4 15409 days
5 15409 days
6 15409 days
```

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

A13. See code below.

```
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()`).
```

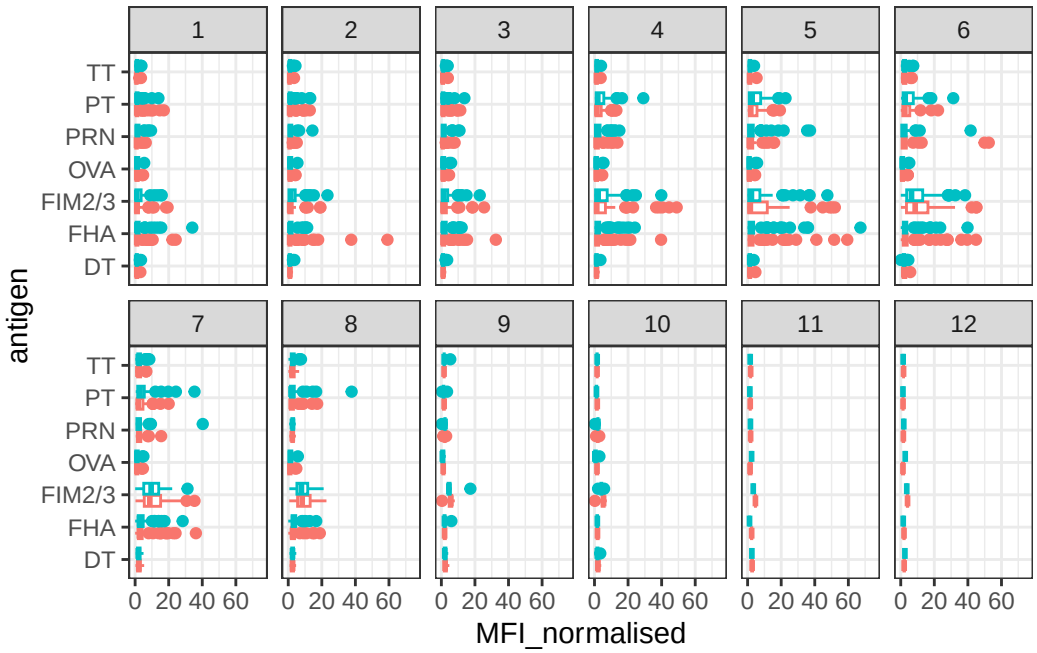


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

A14. The PT and FIM2/3 antigens show increasing levels of IgG antibody titers recognizing them over time. These antigents may have stronger immune responses or longer lasting responses explaining the change over time.

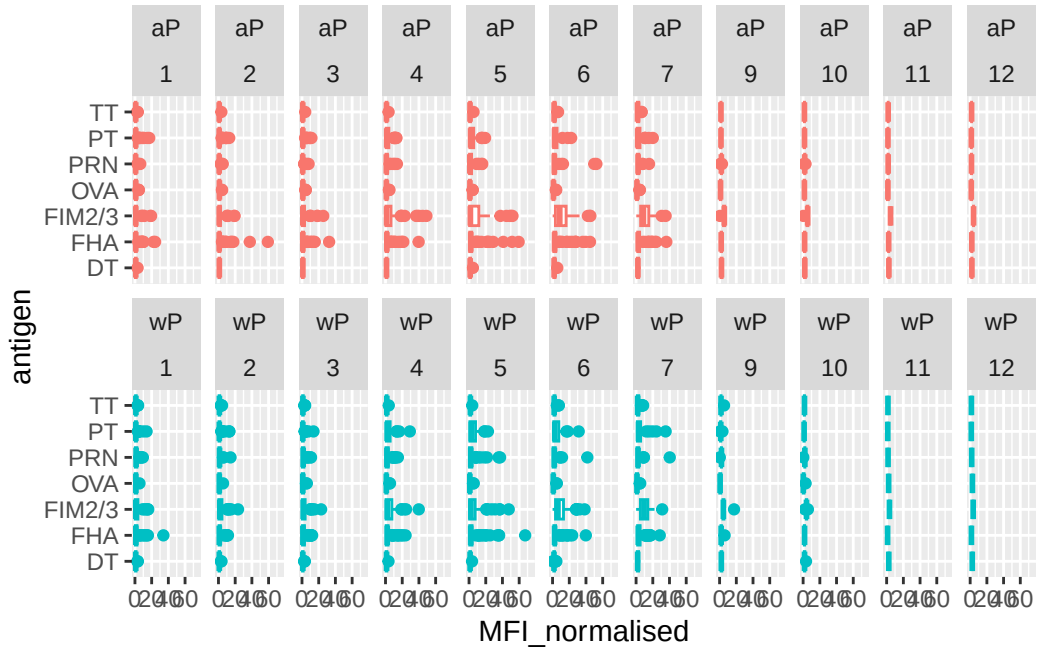
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

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




```
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)
```

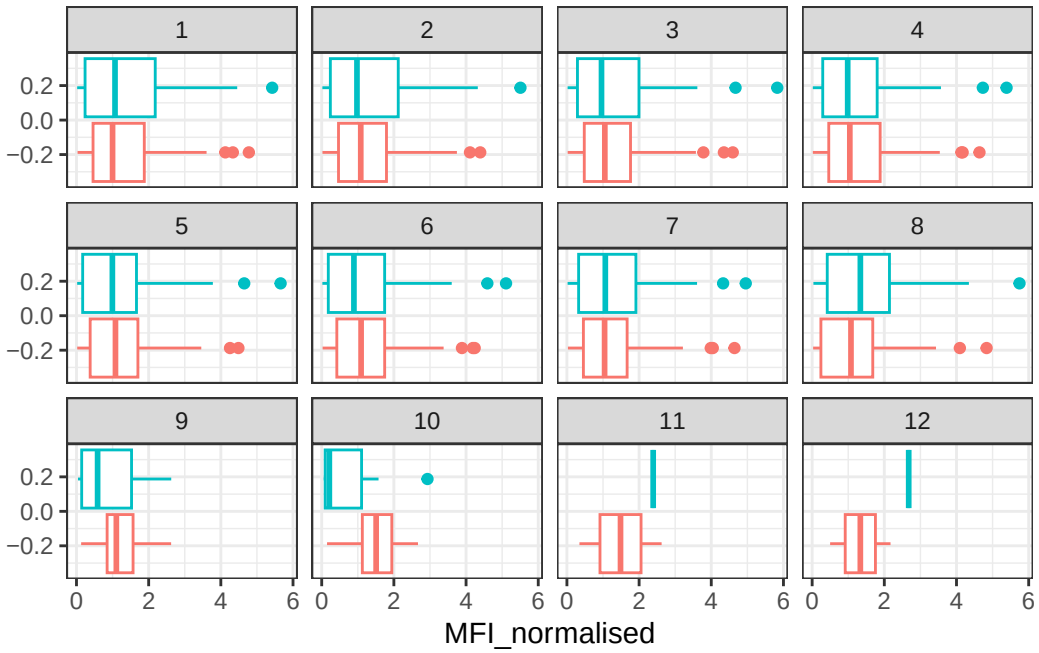
```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



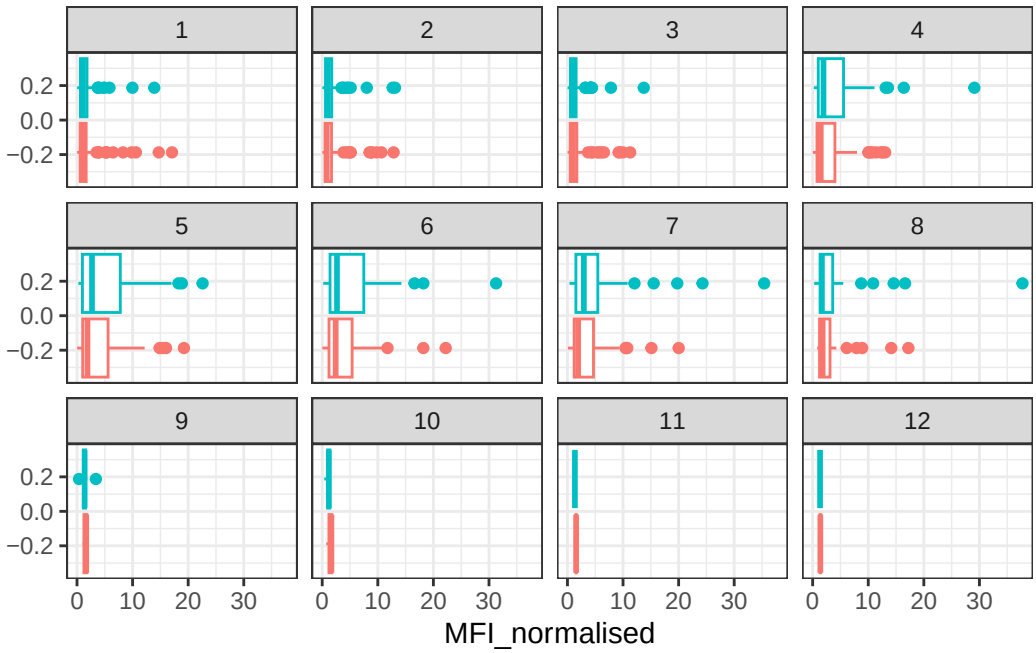
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium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

A15. See below.

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




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



Q16. What do you notice about these two antigens time courses and the PT data in particular?

A16. Keeping the difference in axes in mind, we can see that PT levels rise and exceed OVA values before peaking at visit 5 and declining again. wP (blue) and aP (red) individuals show similar trends in PT levels, but after visit 9 in wP individuals, OVA levels decline while they do not in aP individuals.

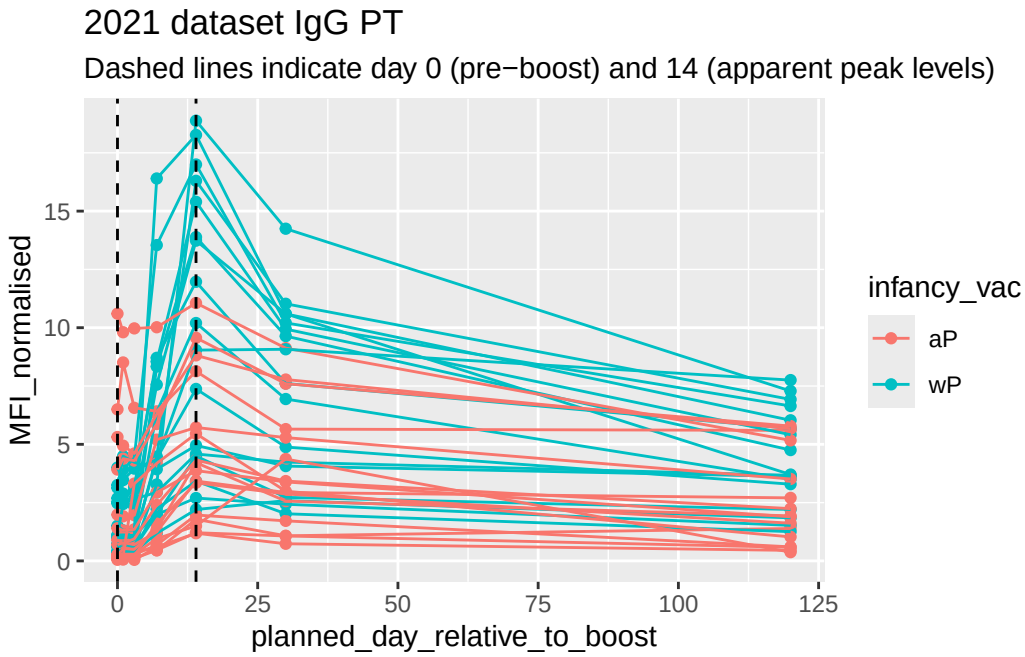
Q17. Do you see any clear difference in aP vs. wP responses?

A17. The wP group experiences higher levels of PT antigen compared to aP and as mentioned in Q16, OVA antigen levels start to decline after visit 9 in the wP group while it remains a constant level after visit 9.

```
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)")
```



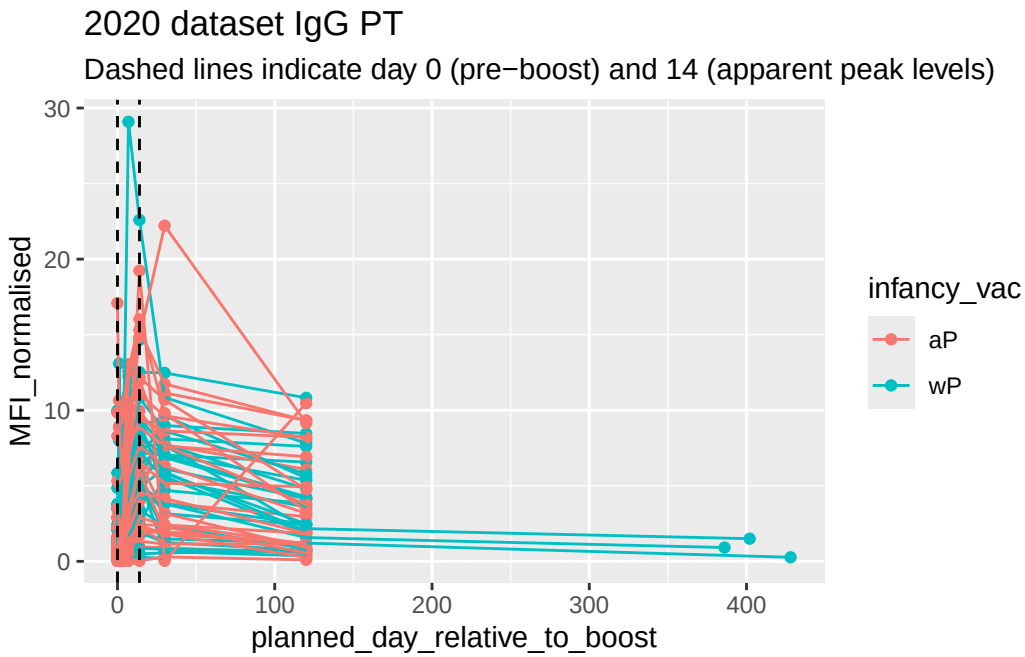
Q18. Does this trend look similar for the 2020 dataset?

A18. Let's look! Not quite, it appears that PT levels in aP individuals showed higher IgG antibody titers recognizing them over time than the 2021 data set.

```
abdata.21 <- abdata %>% filter(dataset == "2020_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="2020 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



5. OBTAINING CMI-PB RNASEQ DATA

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

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

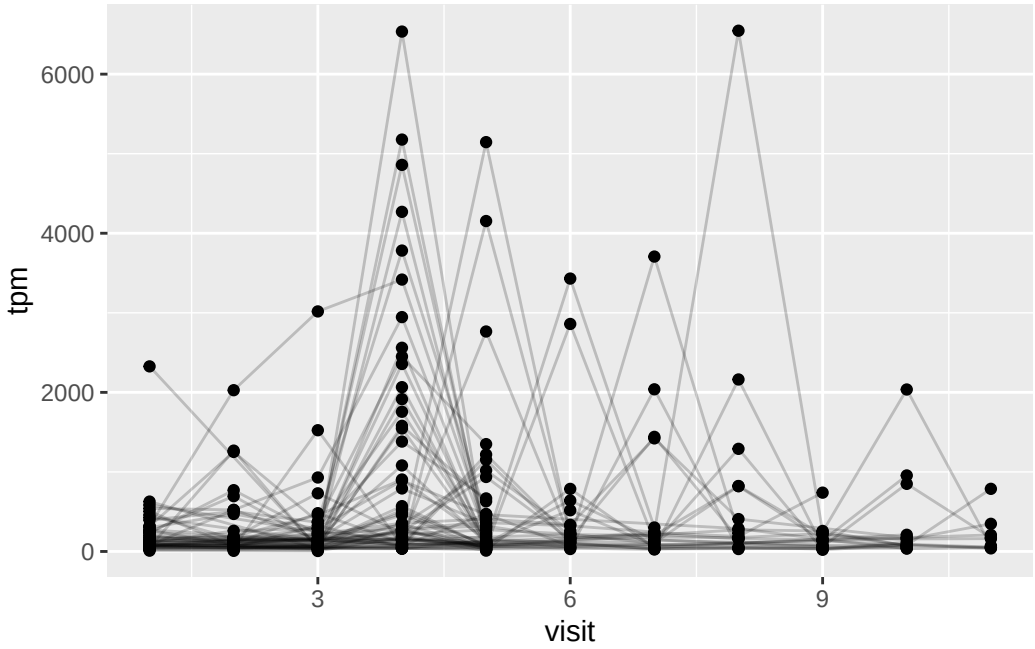
```
#meta <- inner_join(specimen, subject) in console
ssrna <- inner_join(rna, meta)
```

```
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).

A19. See code.

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



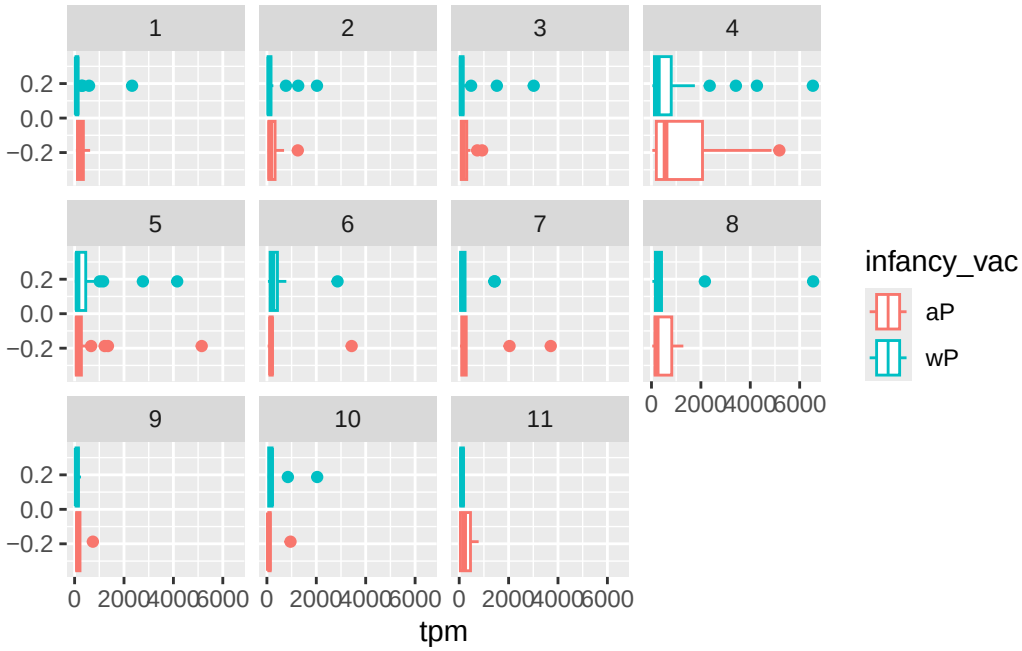
Q20. What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

A20. Expression of this gene is at it's maximum level at visit 4 and then steadly drops in expression afterwards. This is when the body has generated the highest amount of antibodies.

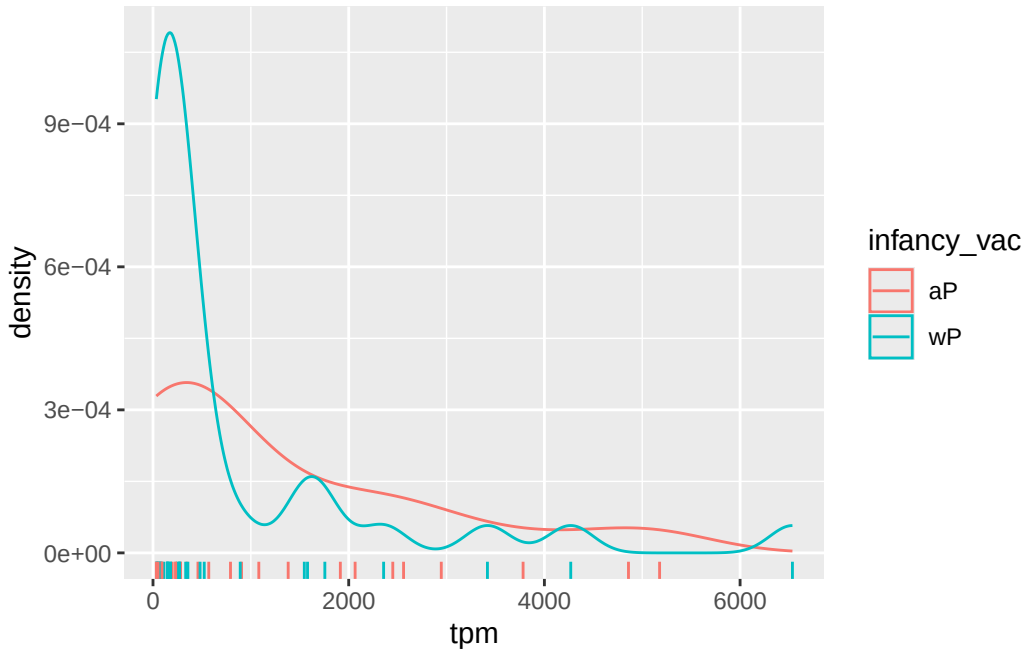
Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

A21. This pattern in time does match the trend of antibody titer data, as the expression of the gene creates the antibodies that we saw eventually decreasing MFI when we analyzed antigen levels per visit.

```
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
9e−04
6e−04
3e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP