Pertussis Mini Project R

Katelyn Brown

3/8/2022

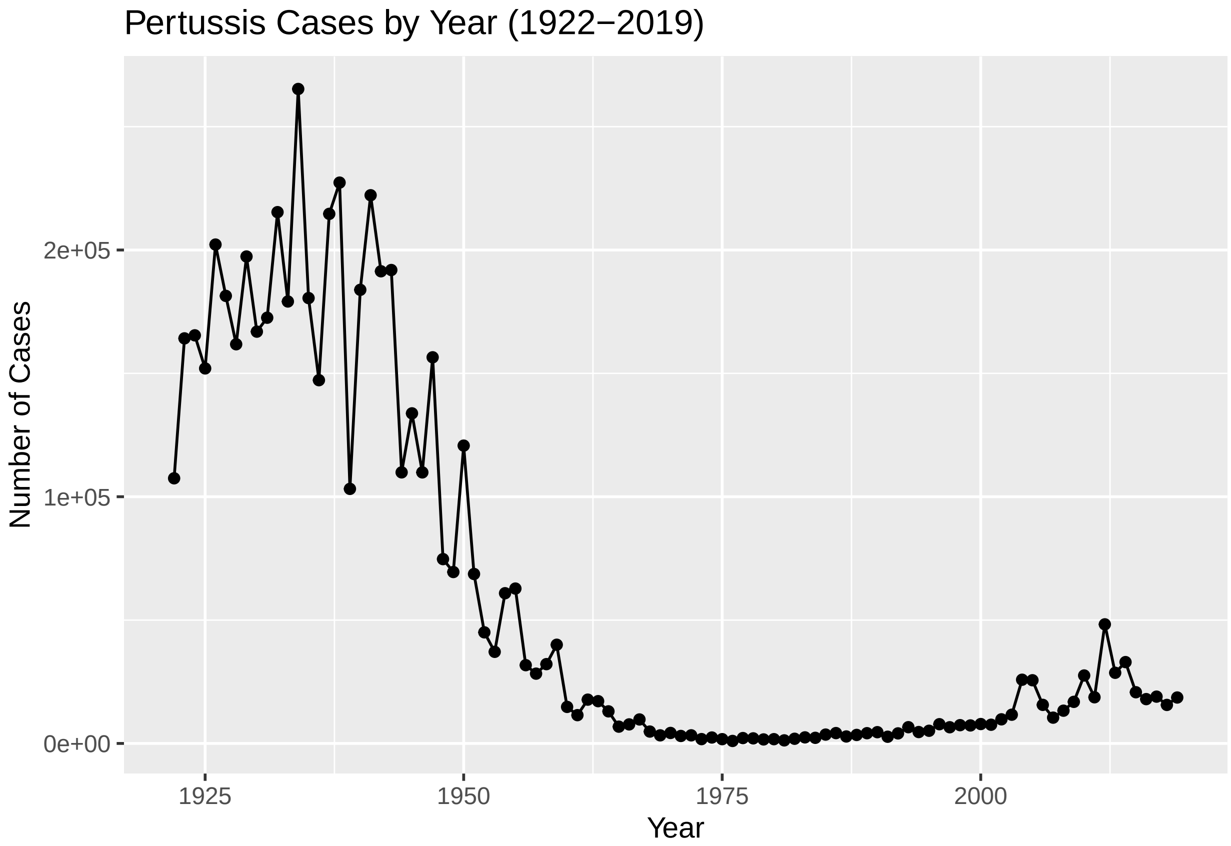
Investigating Pertussis Cases in US by Year

> 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.

```
library(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),
         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)
       )
head(cdc)
```

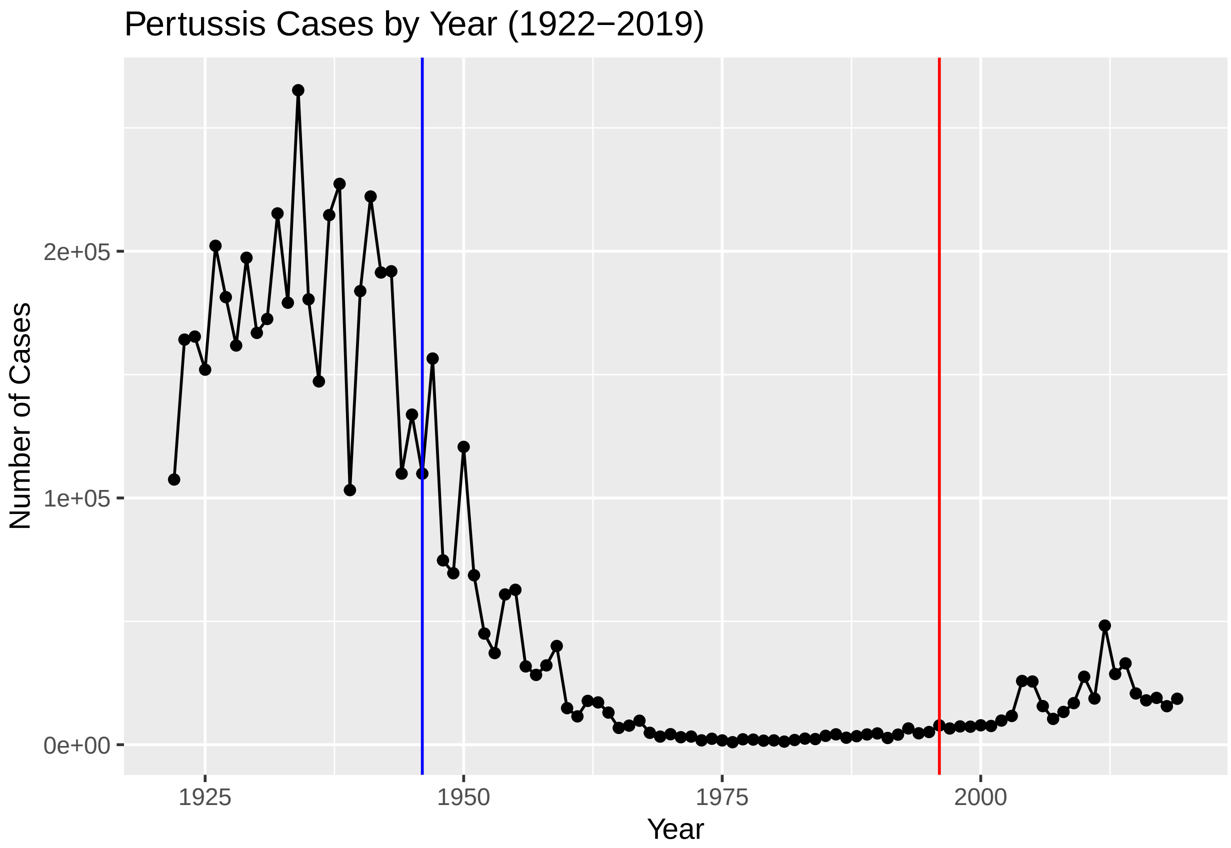
```
##   Year No..Reported.Pertussis.Cases
## 1 1922                       107473
## 2 1923                       164191
## 3 1924                       165418
## 4 1925                       152003
## 5 1926                       202210
## 6 1927                       181411
```

```
# Plot cases
library(ggplot2)
cdcplot <- ggplot(cdc) + aes(Year, No..Reported.Pertussis.Cases) + geom_point() + geom_line() + labs(ti
cdcplot
```



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?

```
cdcplot + geom_vline(aes(xintercept = 1946), color = "blue") + geom_vline(aes(xintercept = 1996), color
```



After the 1946 vaccine rollout, cases of Pertussis dropped dramatically. However, after the 1996 vaccine rollout, cases increased slightly, but has not reached levels similar to those pre-1946.

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

After the introduction of the aP vaccine, cases increased slightly. This trend could be due to vaccine hesitancy, efficacy of the aP vaccine, as well as how the Pertussis bacterium has evolved in response to the vaccine.

Exploring CMI-PB Data

```
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   study_name
## 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 may aP and wP infancy vaccinated subjects are in the dataset?

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

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

There are 47 aP vaccinated and 49 wP vaccinated infancy subjects.

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

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

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

There are 66 female and 30 male patients in this dataset.

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

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

```
##         
##          American Indian/Alaska Native Asian Black or African American
##   Female                             0    18                         2
##   Male                               1     9                         0
##         
##          More Than One Race Native Hawaiian or Other Pacific Islander
##   Female                  8                                         1
##   Male                    2                                         1
##         
##          Unknown or Not Reported White
##   Female                      10    27
##   Male                         4    13
```

The breakdown of race and biological sex is presented in the table.

Working with Dates

> 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?

```r
library(lubridate)
```

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

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

```r
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
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      22      24      25      24      25      26
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      27      31      34      35      39      54
```

The average age of wP individuals is 35 years, the average age of aP individuals is 24. These ages are significantly different, as the difference is 11 years.

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

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

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

```
mean(age_at_boost)
```

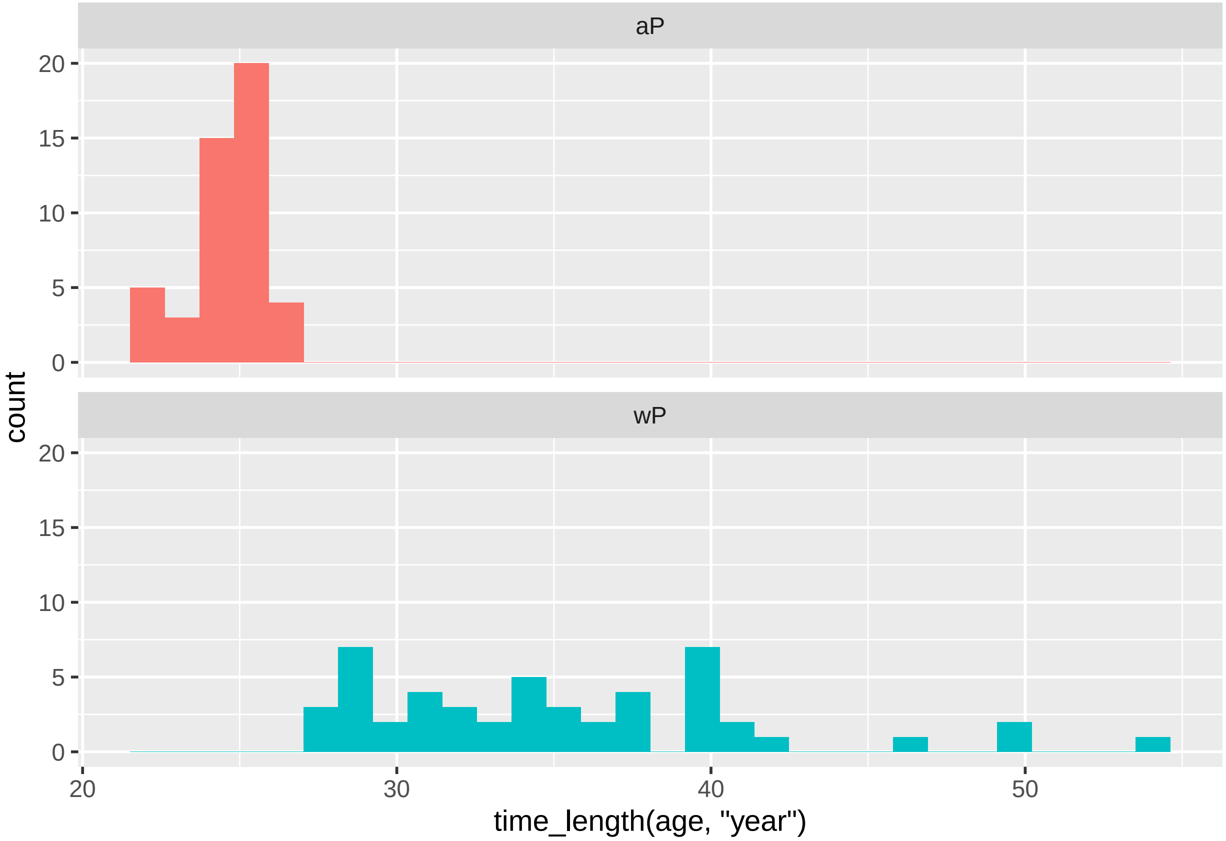
```
## [1] 25.60763
```

The average age of all individuals at time of boost is 25.

> Q9. With the help of a faceted boxplot (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)
```

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



I think these two groups are significantly different when looking at the faceted boxplot. This is because the wP age of boost is significantly higher than the age of boost of the aP group.

Joining Multiple Tables

> 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
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)
head(specimen, 3)
```

```
##   specimen_id subject_id actual_day_relative_to_boost
## 1           1          1                           -3
## 2           2          1                          736
## 3           3          1                            1
##   planned_day_relative_to_boost specimen_type visit
## 1                             0         Blood     1
## 2                           736         Blood    10
## 3                             1         Blood     2
```

```r
head(titer, 3)
```

```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1     IgE               FALSE   Total 1110.21154 UG/ML
## 2           1     IgE               FALSE   Total 2708.91616 IU/ML
## 3           1     IgG                TRUE      PT   68.56614 IU/ML
##   lower_limit_of_detection
## 1                      NaN
## 2                    29.17
## 3                     0.53
```

```r
# Link the tables together
meta <- inner_join(specimen, subject)
```

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

```r
dim(meta)
```

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

```r
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   study_name
## 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 13215 days
## 2 13215 days
## 3 13215 days
## 4 13215 days
## 5 13215 days
## 6 13215 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.

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

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

```
dim(abdata)
```

```
## [1] 32675    20
```

> 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 1413 6141 6141 6141 6141
```

There are 6698 entries for IgE, 1413 for IgG, and 6141 each for IgG1, IgG2, IgG3, and IgG4.

> Q12. What do you notice about the number of visit 8 specimens compared to other visits?

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

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

There is a sharp decrease for visit 8, with only 80 total compared to over 3000 each for the first 1-7 visits. This is because the project is ongoing, and not every patient has had their 8th visit yet.

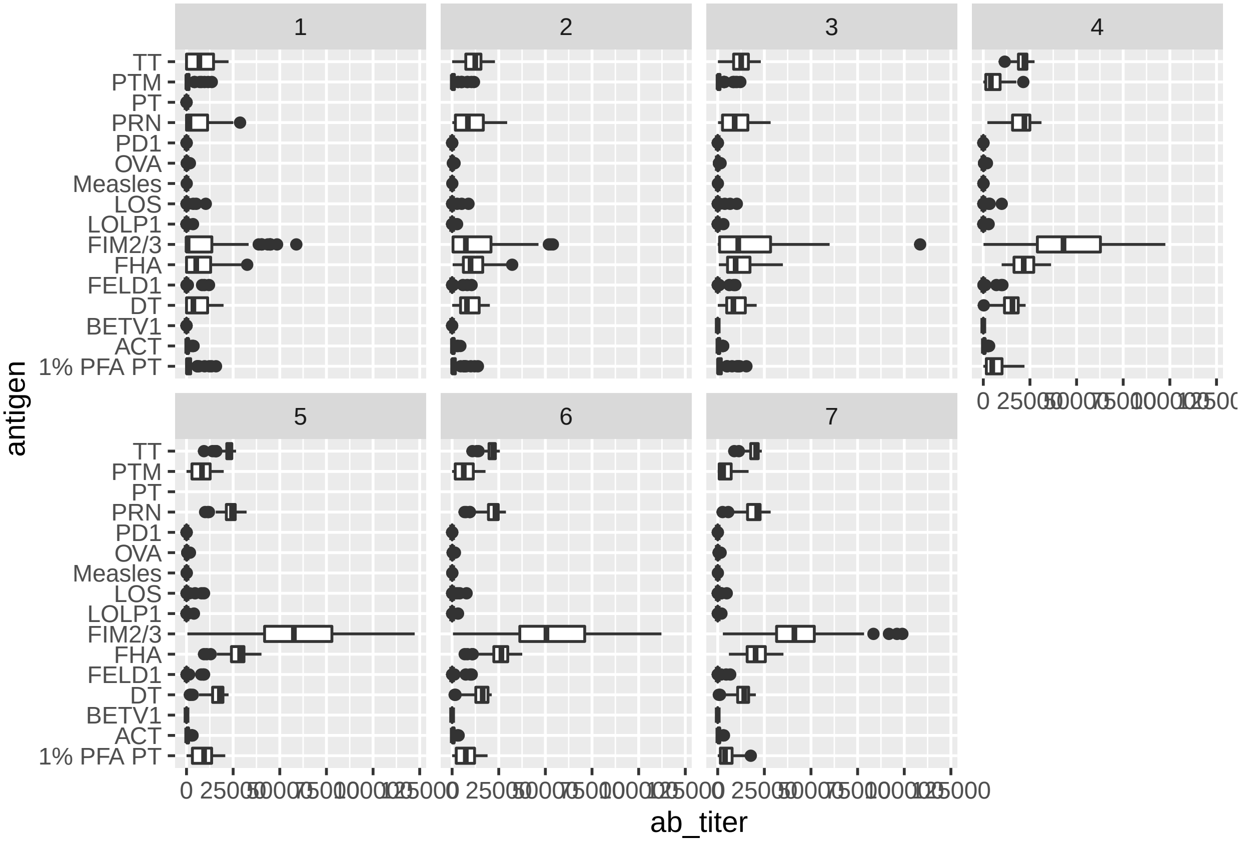
Examining IgG1 Ab titer levels

```
# Exclude visit 8
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1, 3)
```

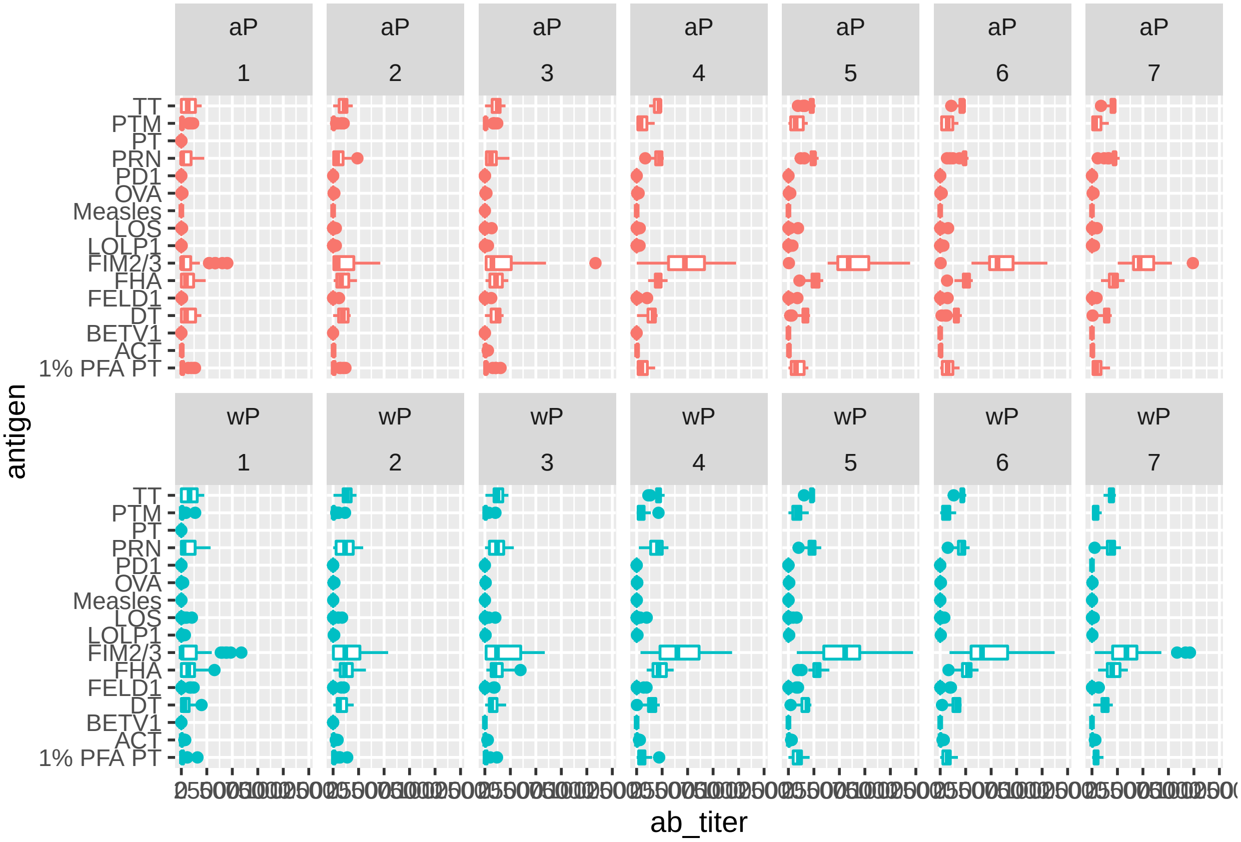
```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1    IgG1                TRUE     ACT 274.355068 IU/ML
## 2           1    IgG1                TRUE     LOS  10.974026 IU/ML
## 3           1    IgG1                TRUE   FELD1   1.448796 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                 3.848750          1                           -3
## 2                 4.357917          1                           -3
## 3                 2.699944          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
##                ethnicity  race year_of_birth date_of_boost   study_name
## 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
##          age
## 1 13215 days
## 2 13215 days
## 3 13215 days
```

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

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



```r
# Edit plot to include aP vs wP
ggplot(ig1) +
  aes(ab_titer, antigen, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

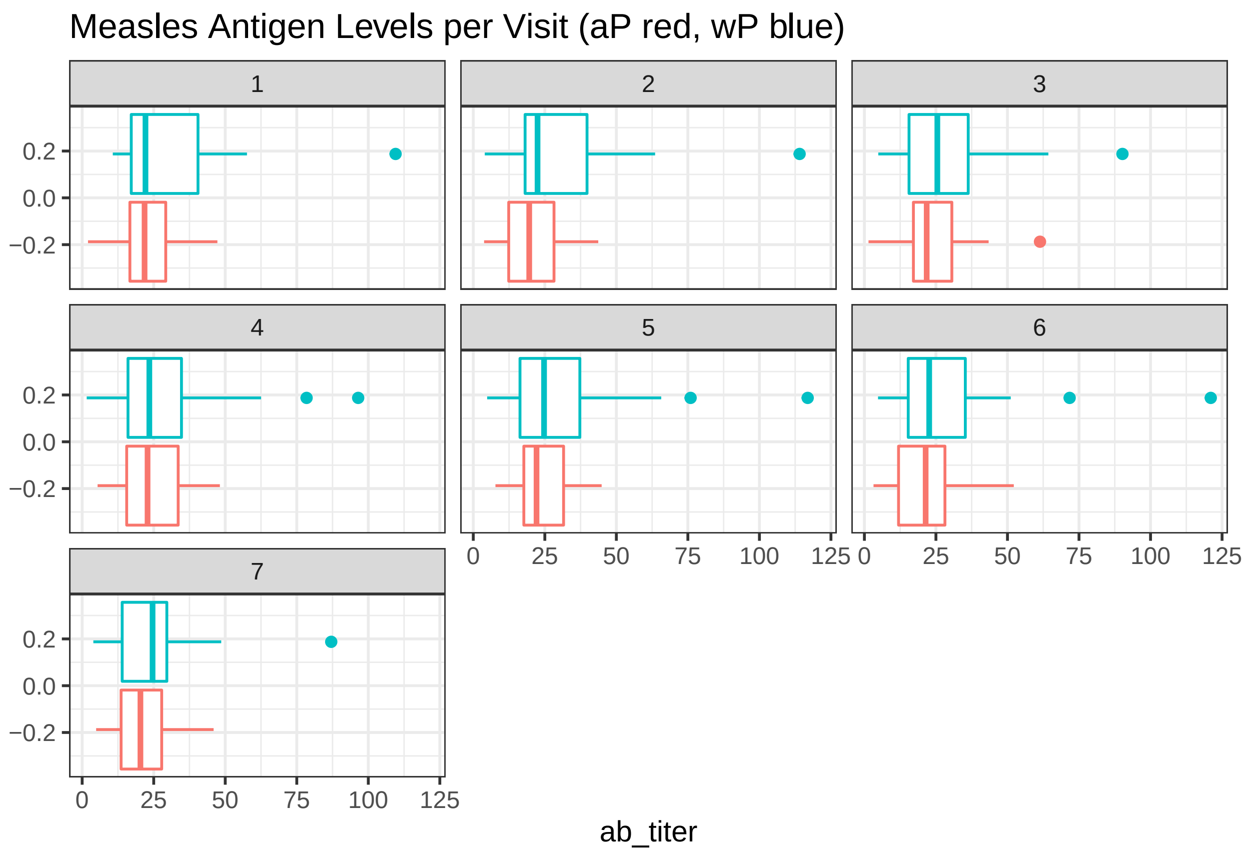


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

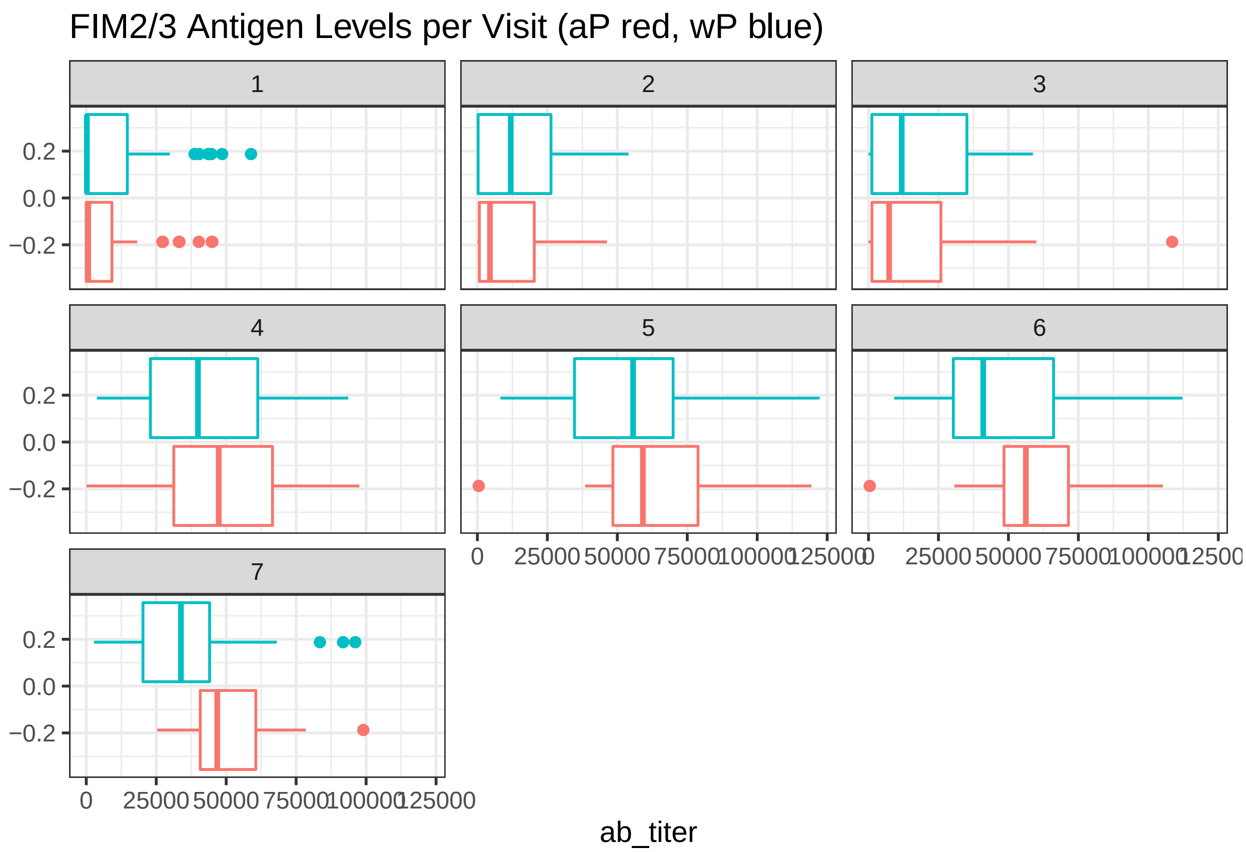
FIM 2/3 shows a large difference in IgG1 antibody level titer over time. PRN and FHA also show differences in IgG1 antibody level titer. The increase in FIM 2/3 antibody titer levels over different visits could indicate that this antigen is present in the vaccine, because at each booster shot visit there is an increase in antibody response to these specific antigens.

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 ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs(title = "Measles Antigen Levels per Visit (aP red, wP blue)")
```



```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs(title = "FIM2/3 Antigen Levels per Visit (aP red, wP blue)")
```



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular?

Based on the Measles antibody response boxplot, the levels of antibody response to measles remain constant compared the FIM2/3 antigen. In the FIM2/3 antigen boxplot, antibody response peaks at around visits 5 and 6, and maintains relatively constant after that. However, there is an increase in antibody responses to the FIM2/3 antigen at every visit leading up to visits 5 and 6.

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

There is a clear difference in aP and wP responses for the FIM2/3 response. Antibody responses for wP remained higher than aP until visit 4, where aP antibody responses then remained higher through visit 7.

Obtaining CMI-PB RNASeq Data

```
# Load data
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"
rna <- read_json(url, simplifyVector = TRUE)
dim(rna)
```

```
## [1] 360   4
```

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

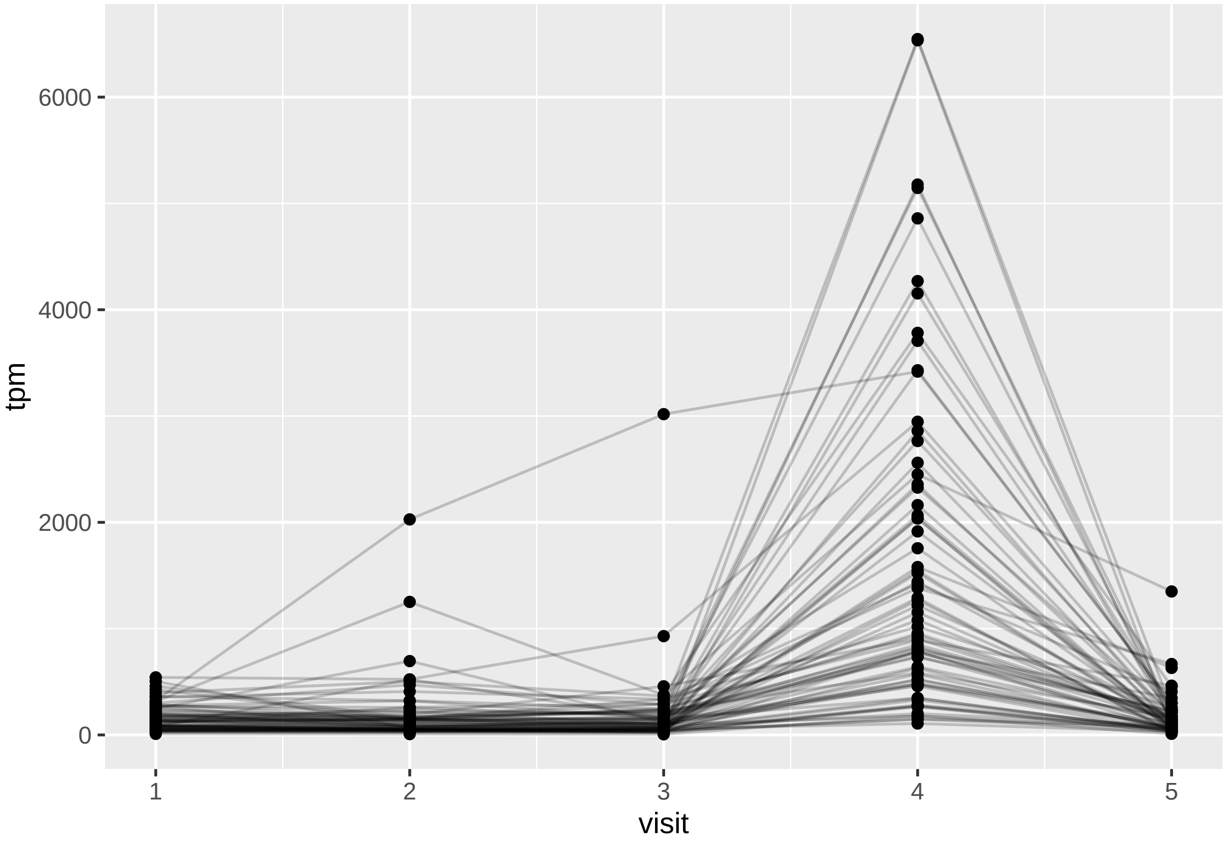
```
## Joining, by = "specimen_id"
```

```
dim(ssrna)
```

```
## [1] 360  17
```

> Q18. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

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



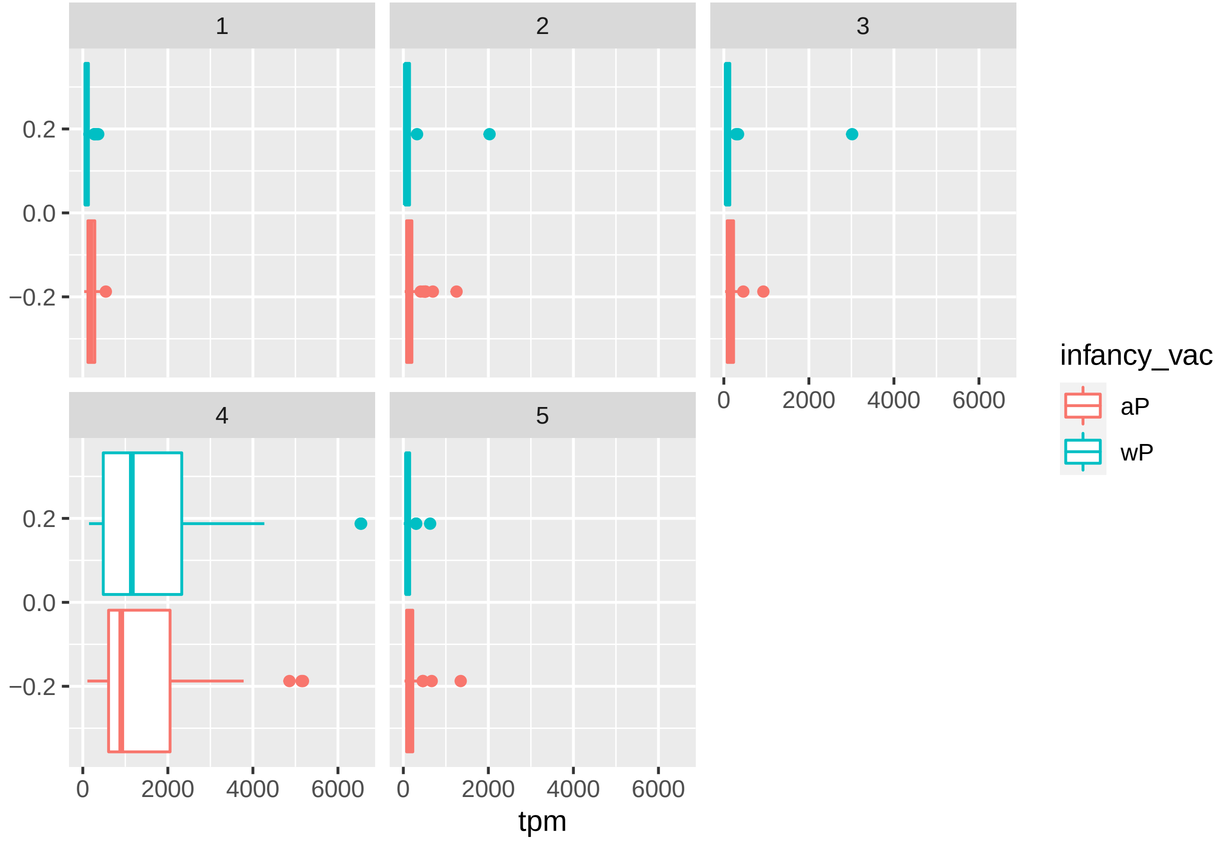
> Q19. What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The expression of the gene is maximized at visit 4.

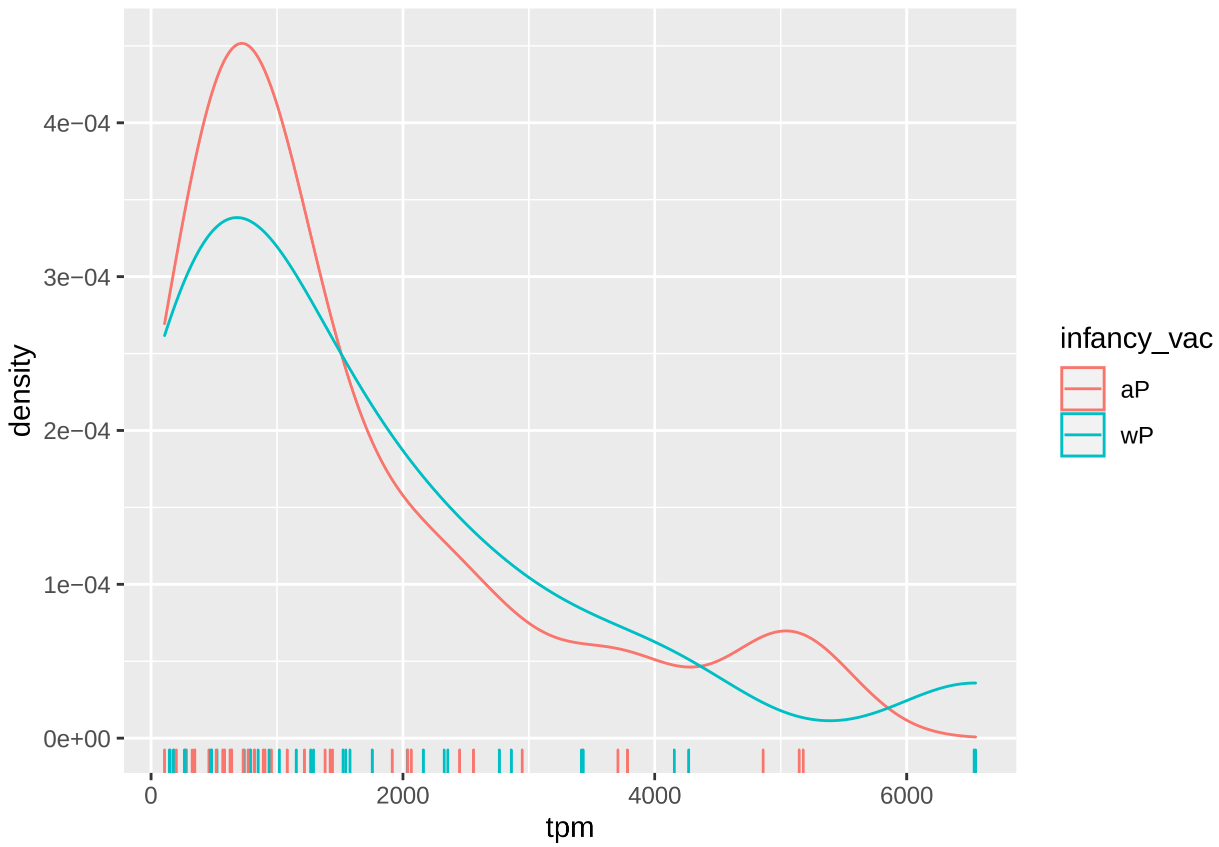
> Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

This pattern does match the trend of the antibody titer data because the boxplots displayed high antibody responses at visit 4, and began to decrease shortly after. There is a slight lag with the boxplot, because the boxplot shows maximum antibody response at visit 5.

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



```
# Focus on specific 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