Class 15: Mini-Project - Investigating Pertussis Resurgence

Taylor F. (A59010460)

3/9/2022

Exploring CDC cases by year data

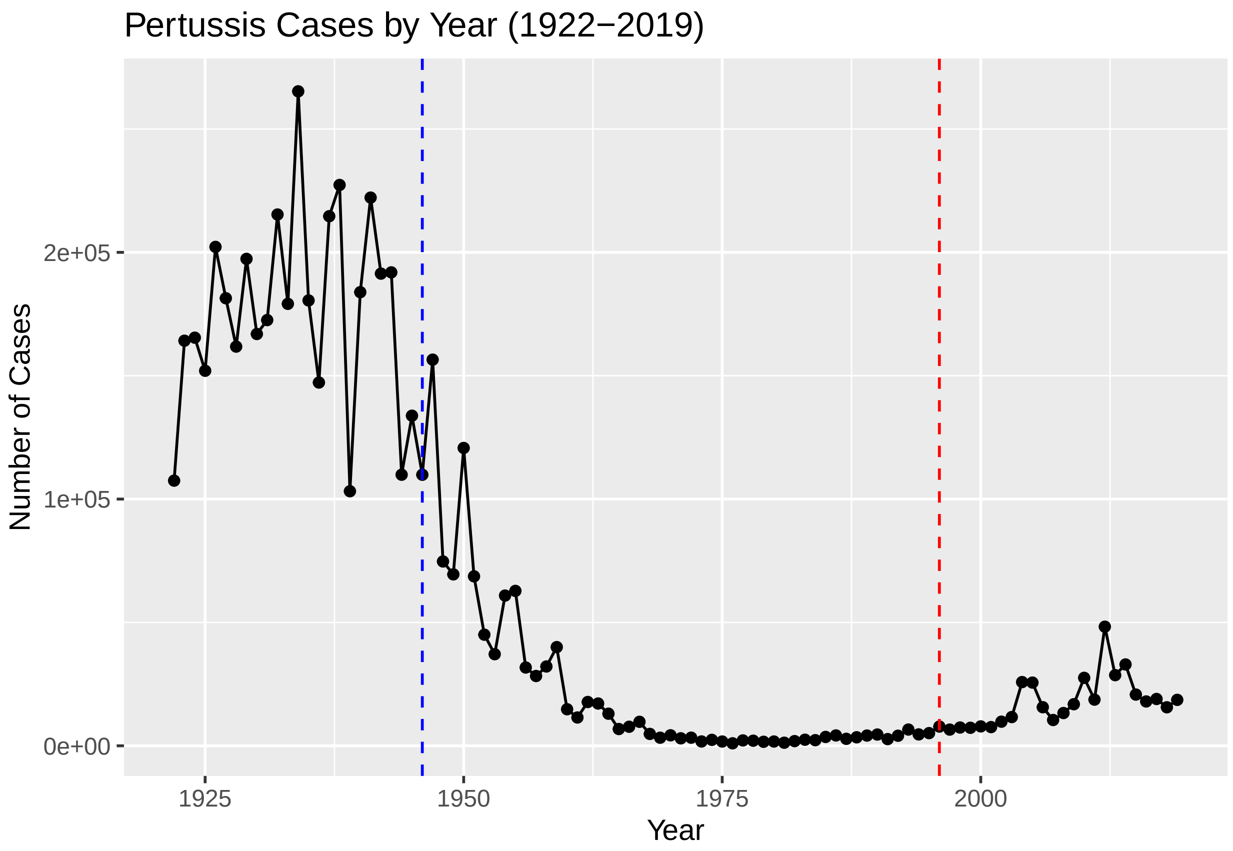
```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
library(datapasta)
library(jsonlite)
#must install.packages("datapasta")
#datapasta lives in the addins dropdown menu now
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)
       )
```

Let's graph this data

```
ggplot(cdc) + aes(Year, No..Reported.Pertussis.Cases) + geom_point() + geom_line() + labs(title = "Pertu
```



Question 3: Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

Pertussis cases are increasing. Some potential explanations include more sensitive testing, vaccination hesitancy, evolution of the pathogen, immunity conferred by the aP vaccine is no longer sufficient as compared to the wP vaccine.

Explore CMI-PB Data

First we need to access JSON format data via the CMI-PB API

```
#Make sure to install jsonlite package (should be done) and 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
```

Question 4: How may aP and wP infancy vaccinated subjects are in the dataset?

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

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

Question 5: How many Male and Female subjects/patients are in the dataset?

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

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

Question 6: What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc…)?

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

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

Joining the Data Sets

First let's complete the APIs

```
# Complete the API URLs...
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)
```

To make sense of this data and learn about wP vs. aP of the Ab titer data, we need to "join" the `subject` with the new tables.

```
#We first need to select the correct join() function (inner, left, right, and full)
#Make sure to load in dplyr so join() can be used
meta <- inner_join(subject, specimen)
```

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

```
dim(meta)
```

```
## [1] 729  13
```

```
#We then need to join the titer dataset with the meta dataset
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
## [1] 32675    19
```

```
#Question 11: 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
```

```
#Question 12: 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 are much fewer visit 8s than the other visit numbers
```

Examining IgG Ab Titer Levels

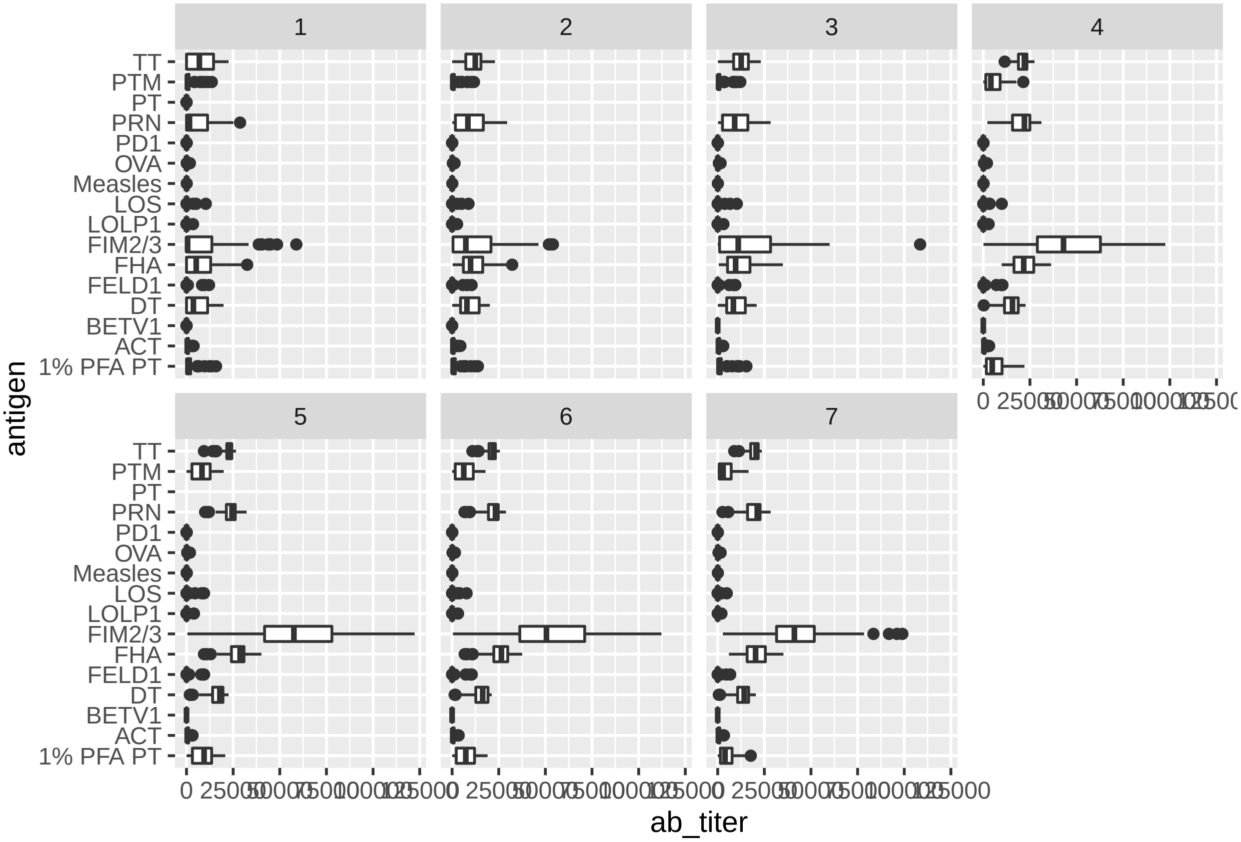
We first want to filter for the IgG isotype and exclude the small number of visit 8 entries

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

```
##   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
## 4           1    IgG1                TRUE   BETV1   0.100000 IU/ML
## 5           1    IgG1                TRUE   LOLP1   0.100000 IU/ML
## 6           1    IgG1                TRUE Measles  36.277417 IU/ML
##   lower_limit_of_detection subject_id infancy_vac biological_sex
## 1                 3.848750          1          wP         Female
## 2                 4.357917          1          wP         Female
## 3                 2.699944          1          wP         Female
## 4                 1.734784          1          wP         Female
## 5                 2.550606          1          wP         Female
## 6                 4.438966          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
## 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
##   actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
## 1                           -3                             0         Blood
## 2                           -3                             0         Blood
## 3                           -3                             0         Blood
## 4                           -3                             0         Blood
## 5                           -3                             0         Blood
## 6                           -3                             0         Blood
##   visit
## 1     1
## 2     1
## 3     1
## 4     1
## 5     1
## 6     1
```

Now let's graph this data

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

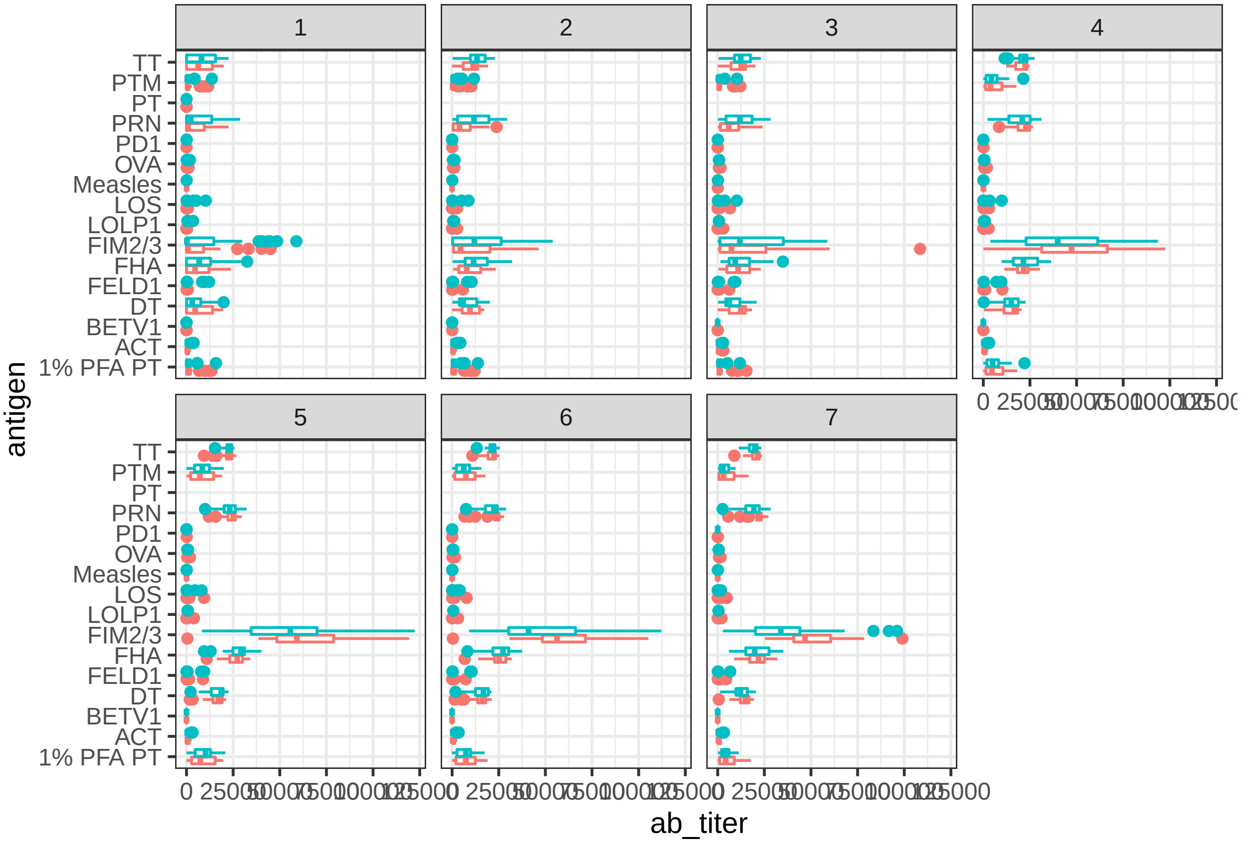


Question 14: What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

It seems as though FIM2/3, FHA, and perhaps PRN are elevated. FIM2/3, FHA, and PRN are pertussis-specific antigens present on the bacterium, so it makes sense they are elevated.

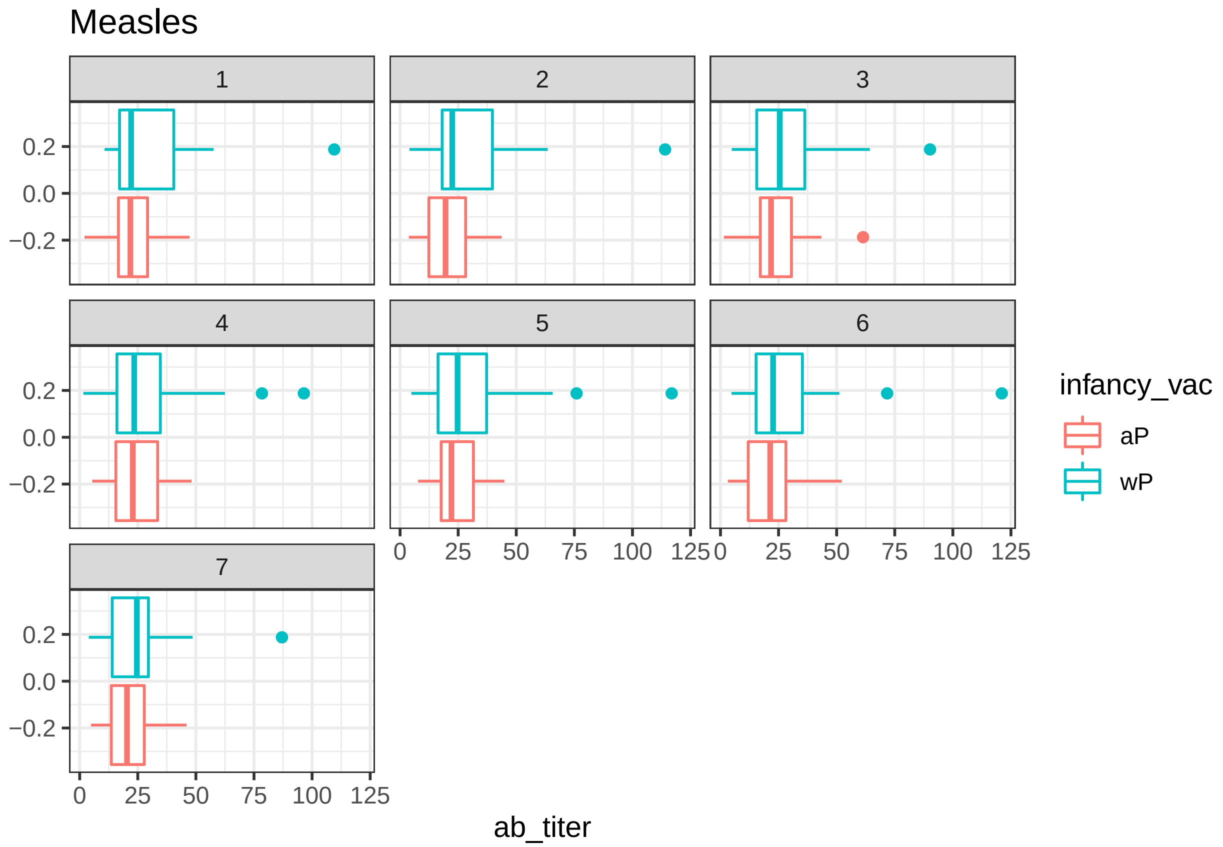
Next, we can differentiate between aP and wP by coloring them differently

```
ggplot(igg1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```

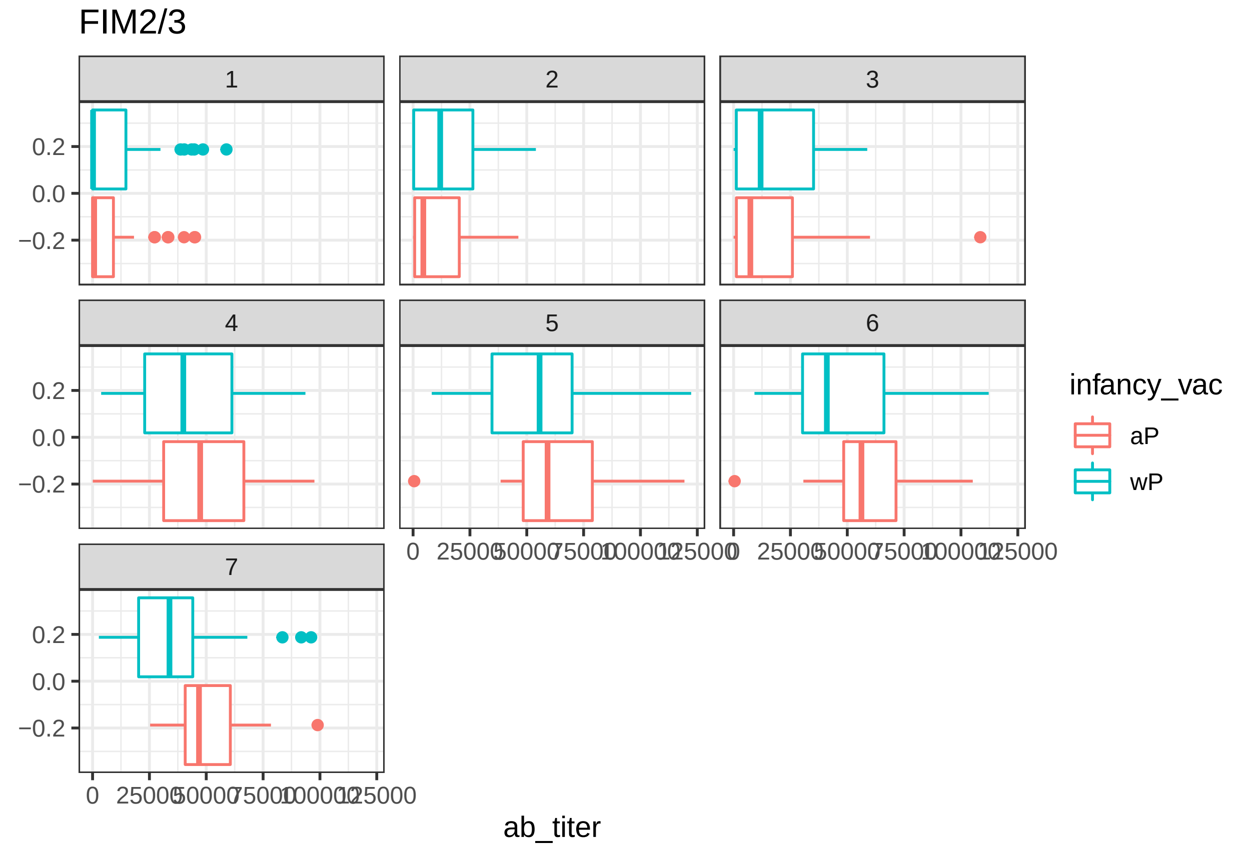


We can also pull out specific antigens to look at their effect on the ab titer over the visits

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



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



Question 16: What do you notice about these two antigens time course and the FIM2/3 data in particular?

FIM2/3 responds more robustly over time than measles, and seems to reach its maximum effect around visit 5, decreasing a bit after that.

Question 17: Do you see any clear difference in aP vs. wP responses?

It seems as though aP peaks higher and fades more slowly than wP, but this data alone is not enought to make any definitive statements about differences between the two.

Obtaining CMI-PB RNASeq data

Let's read available RNA-Seq data for the IGHG1 gene into R and investigate the time course of it's gene expression values.

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

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

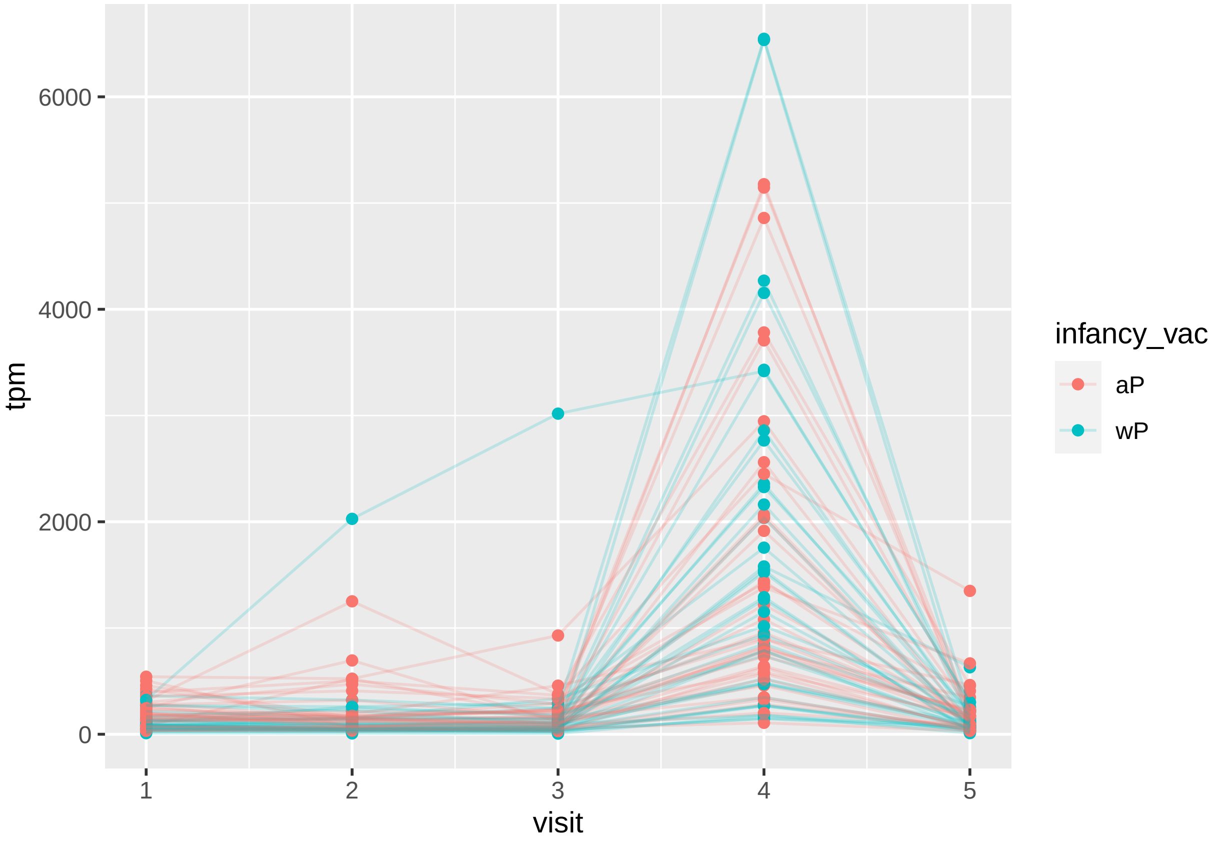
We now need to join the rna dataset with the meta dataset, which itself is a joining of the subject and speciment datasets. This will allow us to look at this genes TPM expression values over aP/wP status and at different visits.

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

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

We can now plot this data

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



Question 19: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

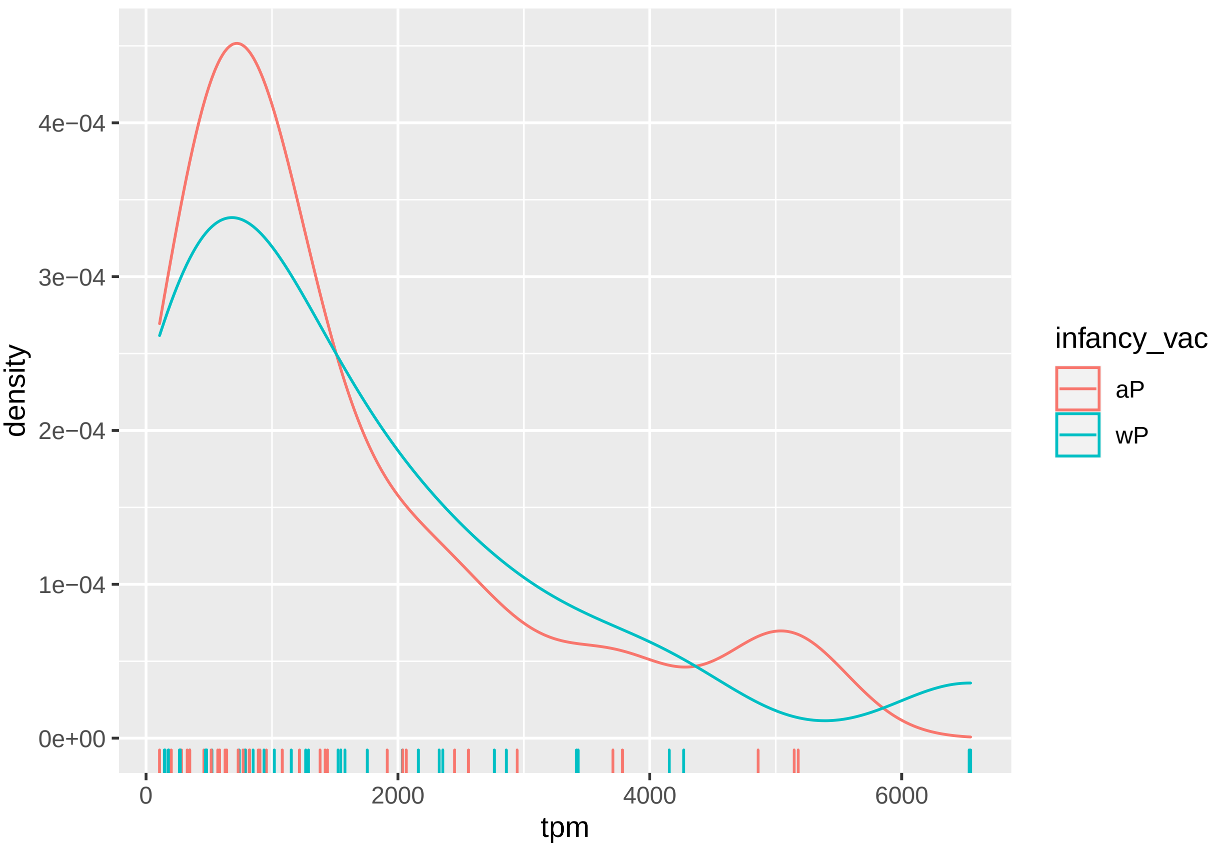
The expression of this gene peaks extremely rapidly at visit 4 and drops precipitously by the next visit. But when it is at its maximum, it is quite variable.

Question 20: Does this pattern in time match the trend of antibody titer data? If not, why not?

No, not necessarily. This data indicates that much IgG is made around visit 4. Antibodies are long-lived, and as the antibody titer shows, much of that IgG made on visit 4 is still around for the next visit, but is not necessarily being synthesized anymore (according to this graph).

With all of this in mind, we do not see any differences between aP and wP as it relates to differential IGHG1 expression (even during the most extreme 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