lab15r

Rachel Kraft

2022-03-08

Section 1: Investigating Pertussis by year in the US

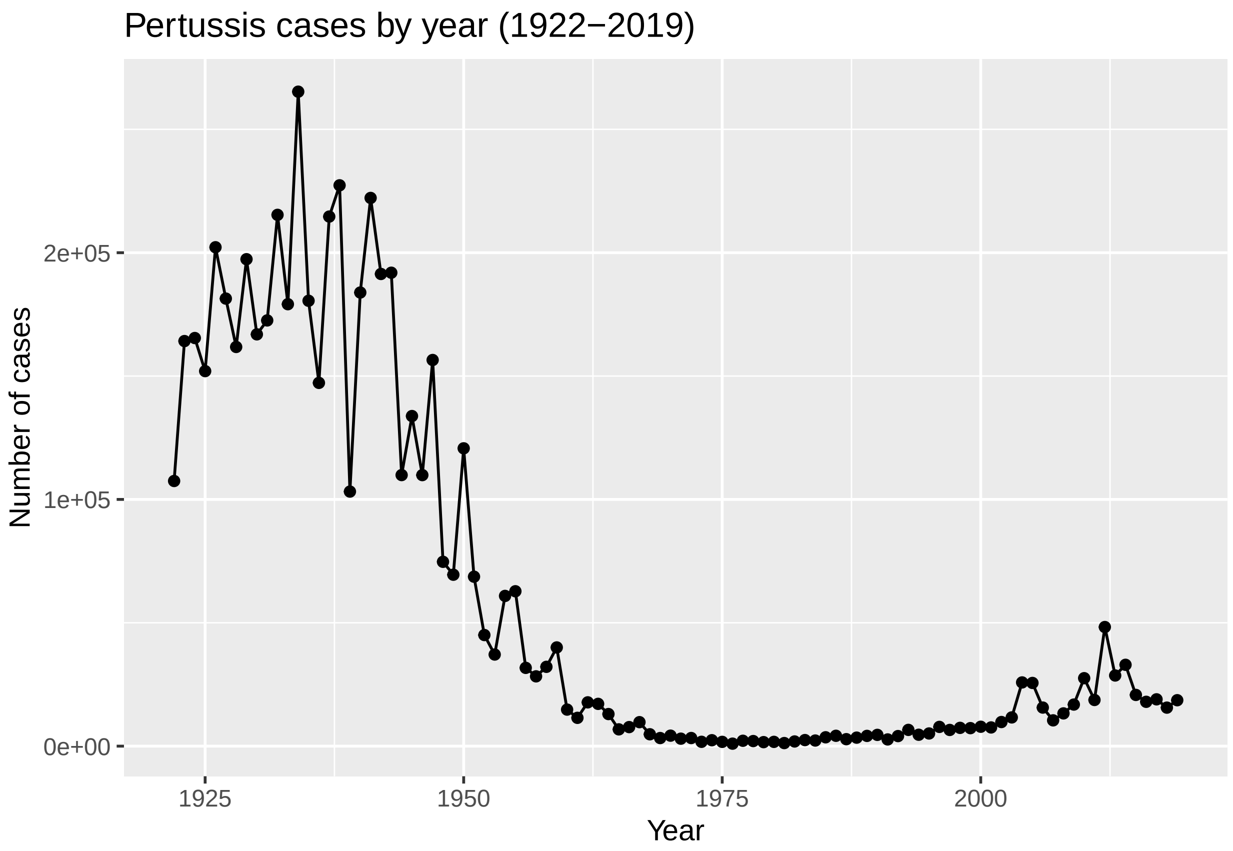
Q1. With the help of the R “addin” package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

install.packages(“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)
       )
```

```
library(tidyverse)
```

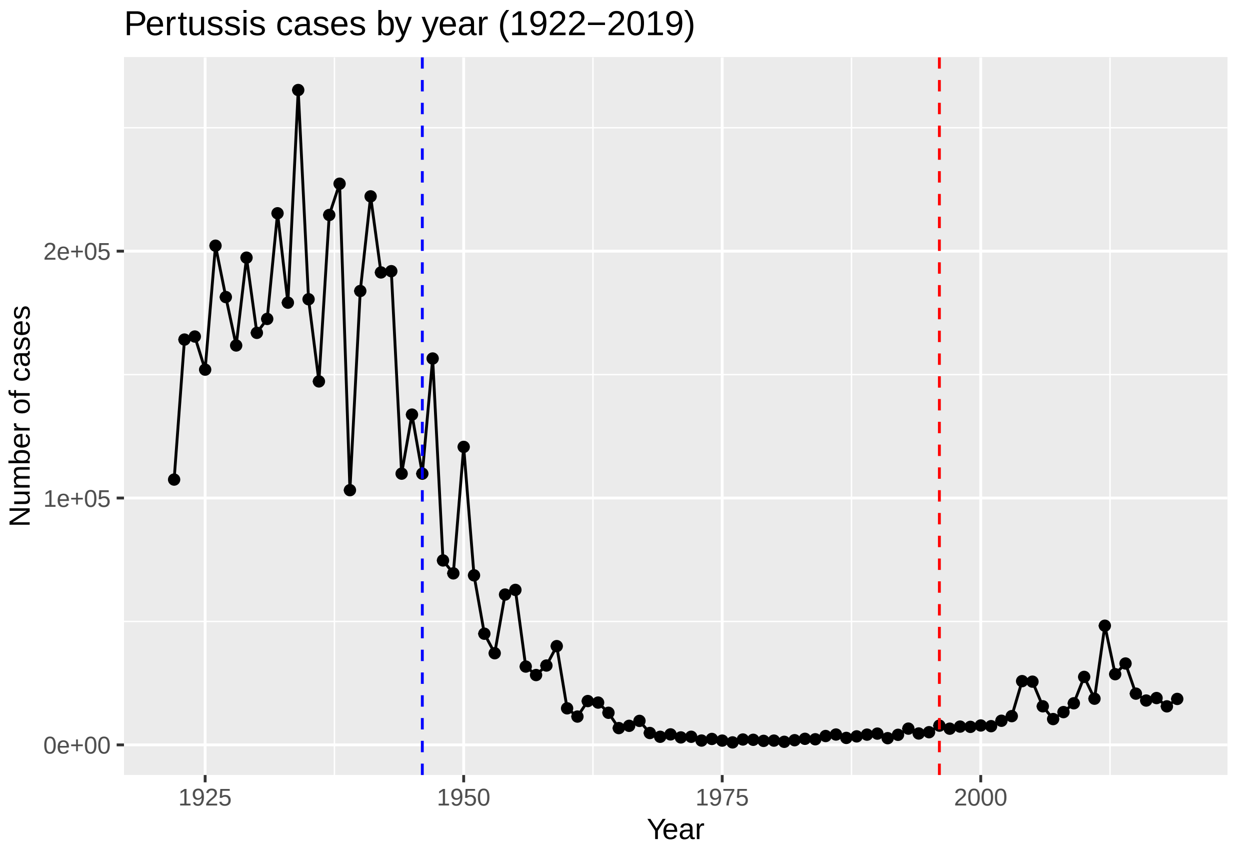
```
ggplot(cdc) +
  aes(x=Year, y=No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(y="Number of cases", title="Pertussis cases by year (1922-2019)")
```



Section 2: Tale of two vaccines

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?

```
ggplot(cdc) +
  aes(x=Year, y=No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(y="Number of cases", title="Pertussis cases by year (1922-2019)")+
  geom_vline(xintercept=1946, color="blue", linetype=2) +
  geom_vline(xintercept=1996, color="red", linetype=2)
```



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

Since the aP vaccine was new, there was possibly more hesitancy in getting the vaccine leading to an increase in observed cases. Also, there could have been more testing that was more accurate, or the bacteria could have mutated and evolved to be more resistant to the vaccines. Another possible reason for the increase in cases is that people who were vaccinated with the old vaccine might have decreased immunity later in life.

Section 3: Exploring CMI-PB data

Why is this vaccine preventable disease rising? Use jsonlite package to read JSON data, made of key-value pairs where a key is associated with a certain value.

install.packages("jsonlite")

```
library(jsonlite)
```

Let's read the main subject database table from the CMI-PB API

```
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
```

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

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

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

> Q6. 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
```

> working with dates

```
library(lubridate)
```

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

we can use the ymd() function to tell lubridate the format of our particular date and then the time_length() function to convert days to years

> 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?

```
subject$age <- today() - ymd(subject$year_of_birth)
time_length(subject$age, "year")
```

```
##  [1] 36.18070 54.18207 39.18138 34.18207 31.18138 34.18207 41.18001 37.18001
##  [9] 26.18207 40.18070 36.18070 40.18070 25.18001 29.18001 33.18001 35.18138
## [17] 42.18207 25.18001 28.18070 35.18138 29.18001 27.18138 29.18001 32.18070
## [25] 46.18207 50.18207 50.18207 32.18070 24.18070 24.18070 31.18138 27.18138
## [33] 27.18138 24.18070 24.18070 34.18207 29.18001 35.18138 30.18207 29.18001
## [41] 24.18070 23.18138 25.18001 22.18207 24.18070 22.18207 22.18207 25.18001
## [49] 23.18138 24.18070 22.18207 26.18207 23.18138 24.18070 22.18207 41.18001
## [57] 39.18138 37.18001 31.18138 30.18207 34.18207 39.18138 25.18001 40.18070
## [65] 25.18001 34.18207 33.18001 25.18001 32.18070 39.18138 31.18138 25.18001
## [73] 24.18070 25.18001 37.18001 28.18070 37.18001 25.18001 24.18070 24.18070
## [81] 25.18001 24.18070 26.18207 24.18070 25.18001 25.18001 25.18001 24.18070
## [89] 24.18070 25.18001 25.18001 25.18001 26.18207 25.18001 25.18001 25.18001
```

```
library(dplyr)
```

```
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
```

```
# wP
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
```

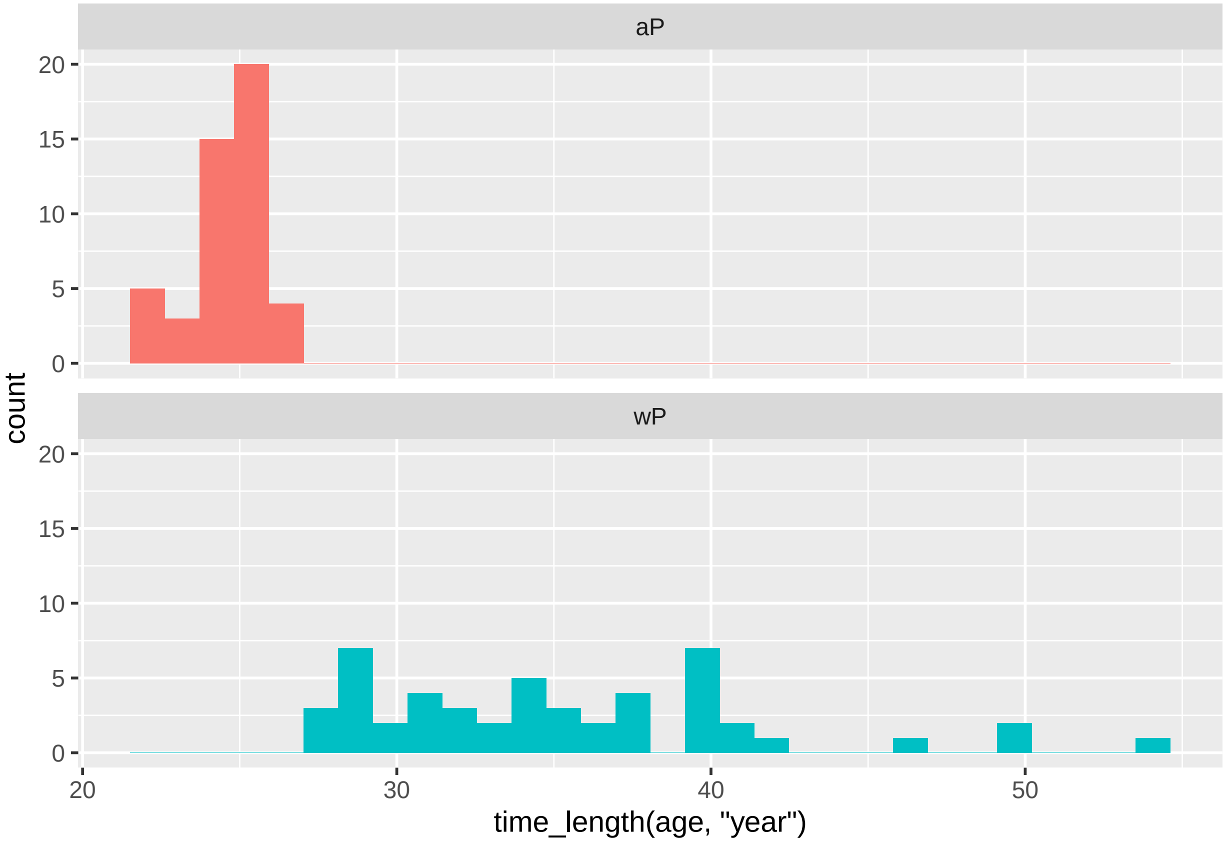
The median age of wP individuals is around 10 years higher than aP individuals

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

> 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‘.
```



Read the specimen and ab_titer tables into R and store the data as specimen and titer named data frames.

```
# 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 know whether a given specimen_id comes from an aP or wP individual we need to link (a.k.a. “join” or merge) our specimen and subject data frames

> 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

```
meta <- inner_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   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?

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

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

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

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

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

The visit 8 specimens are significantly lower than the others, thousands less

Section 4: examine IgG1 Ab titer levels

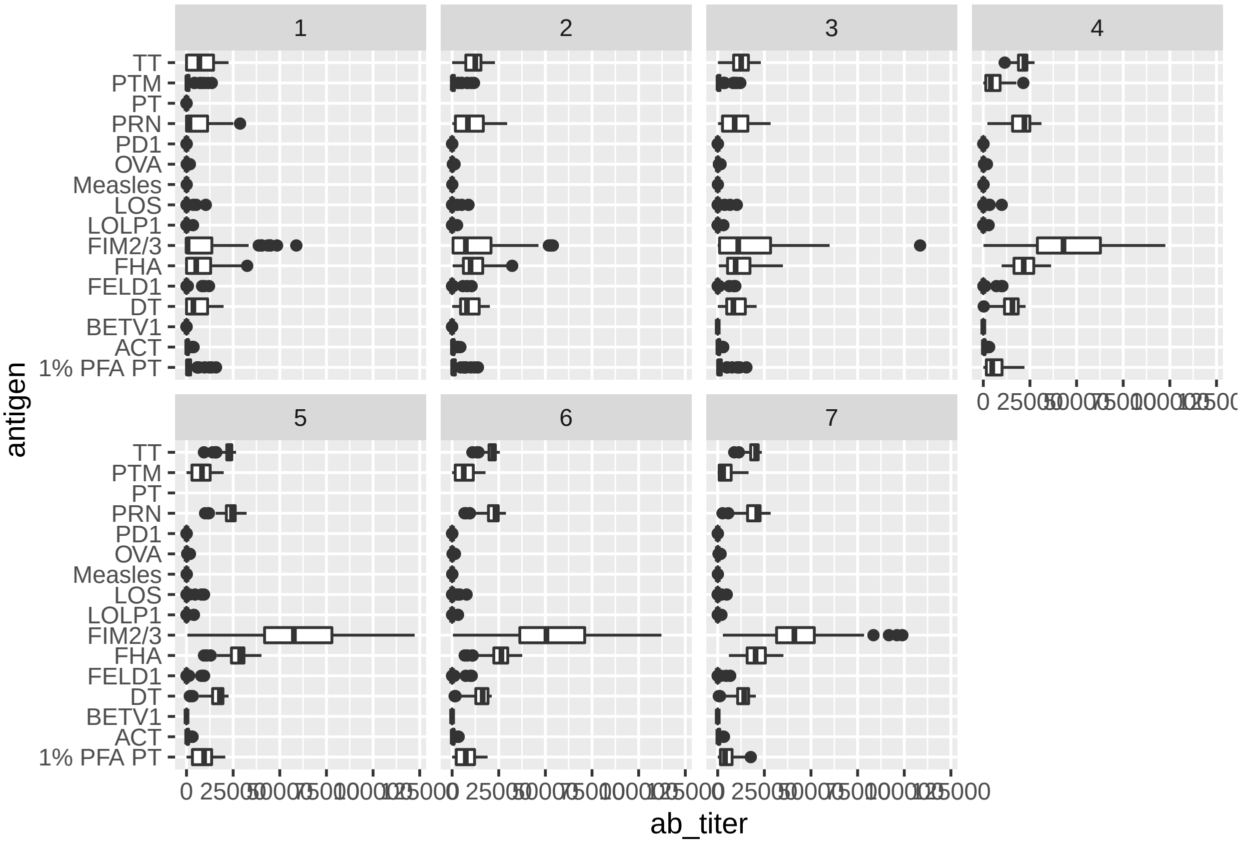
Now using our joined/merged/linked abdata dataset filter() for IgG1 isotype and exclude the small number of visit 8 entries

```r
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 actual_day_relative_to_boost
## 1                 3.848750          1                           -3
## 2                 4.357917          1                           -3
## 3                 2.699944          1                           -3
## 4                 1.734784          1                           -3
## 5                 2.550606          1                           -3
## 6                 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   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
```

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

```
ggplot(igg1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```

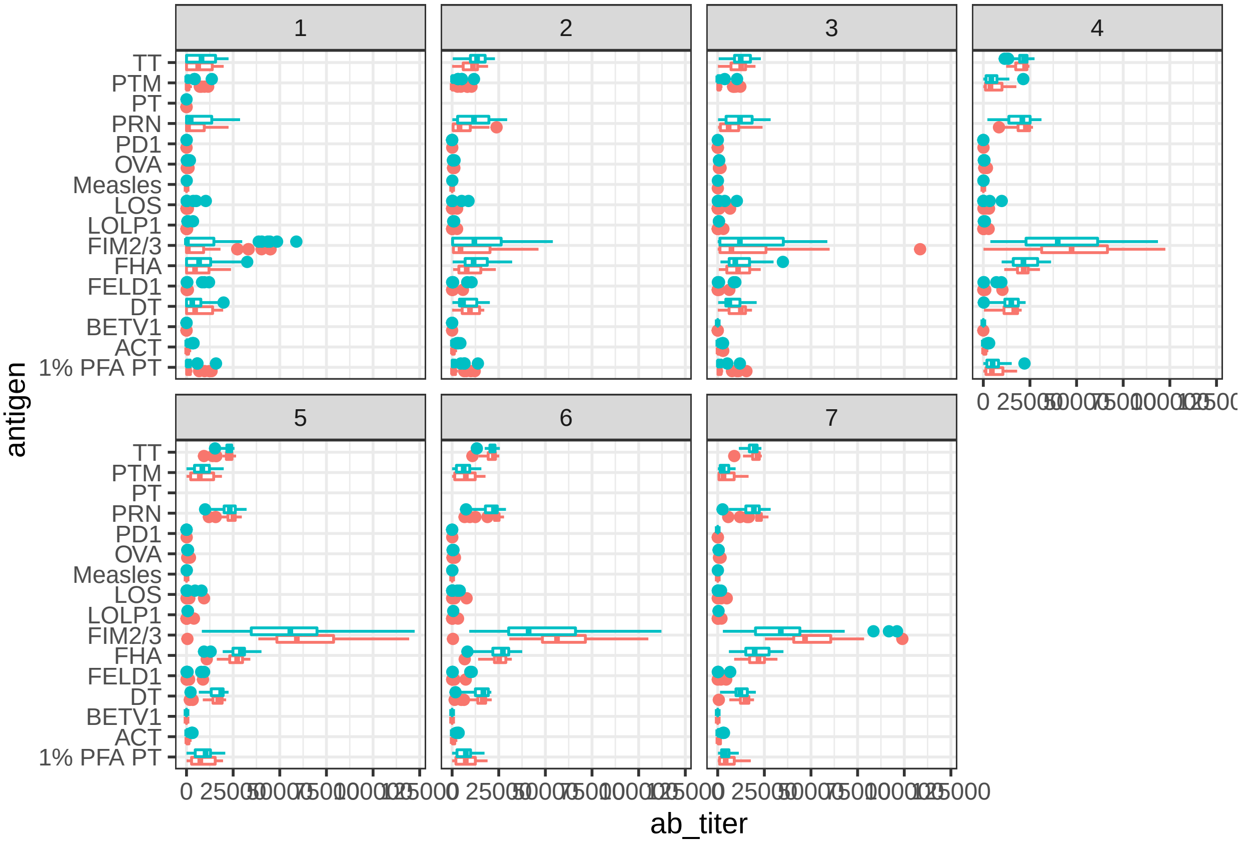


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3, PRN, TT, FHA, DT; These are the antigens that respond to the pertussis bacteria

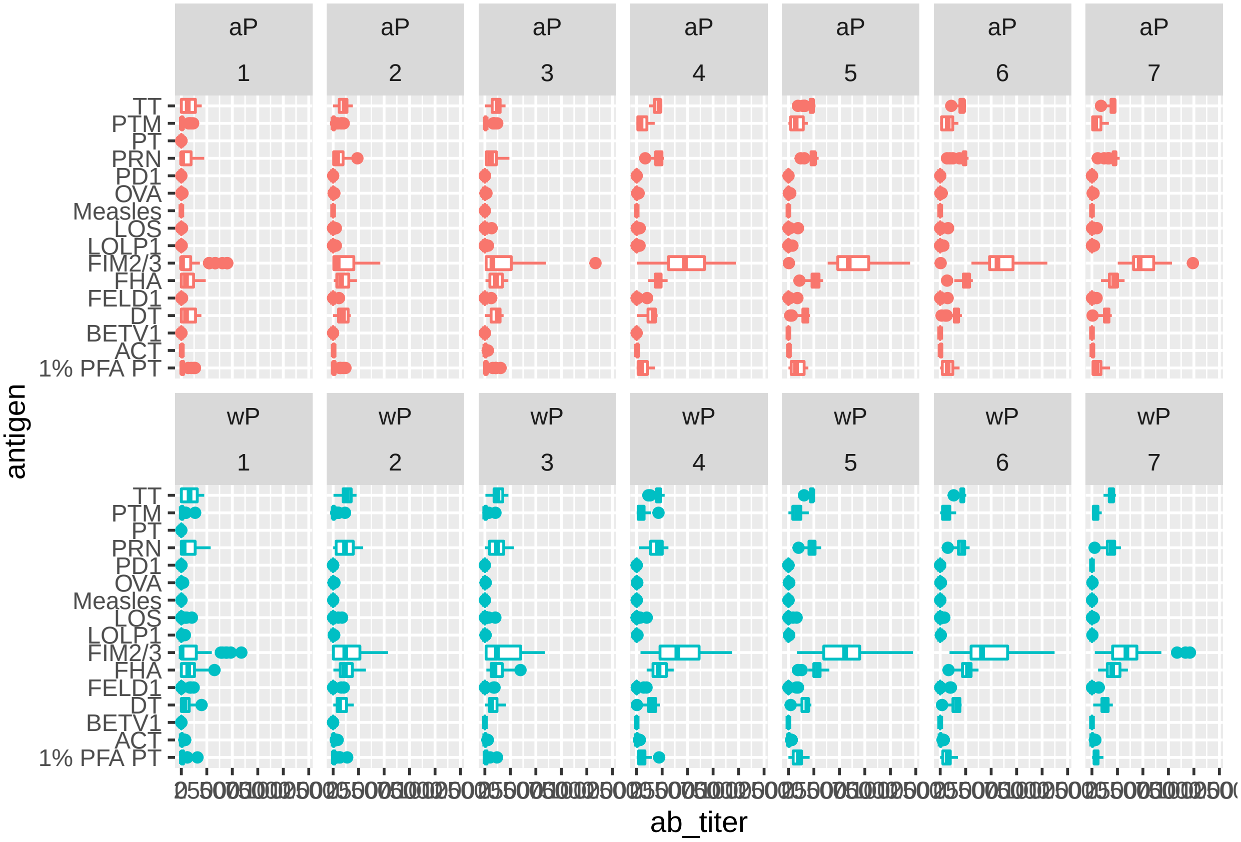
We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status

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



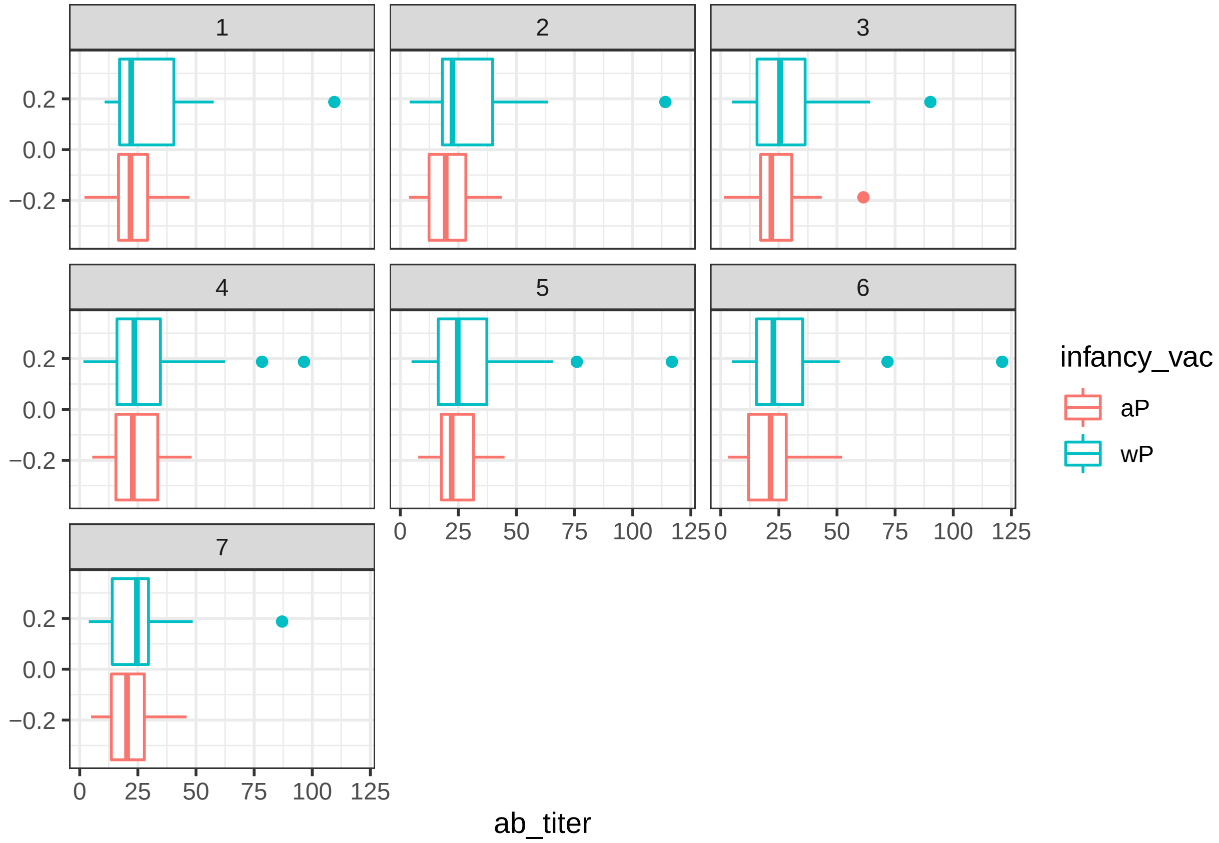
add infancy_vac to the faceting

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



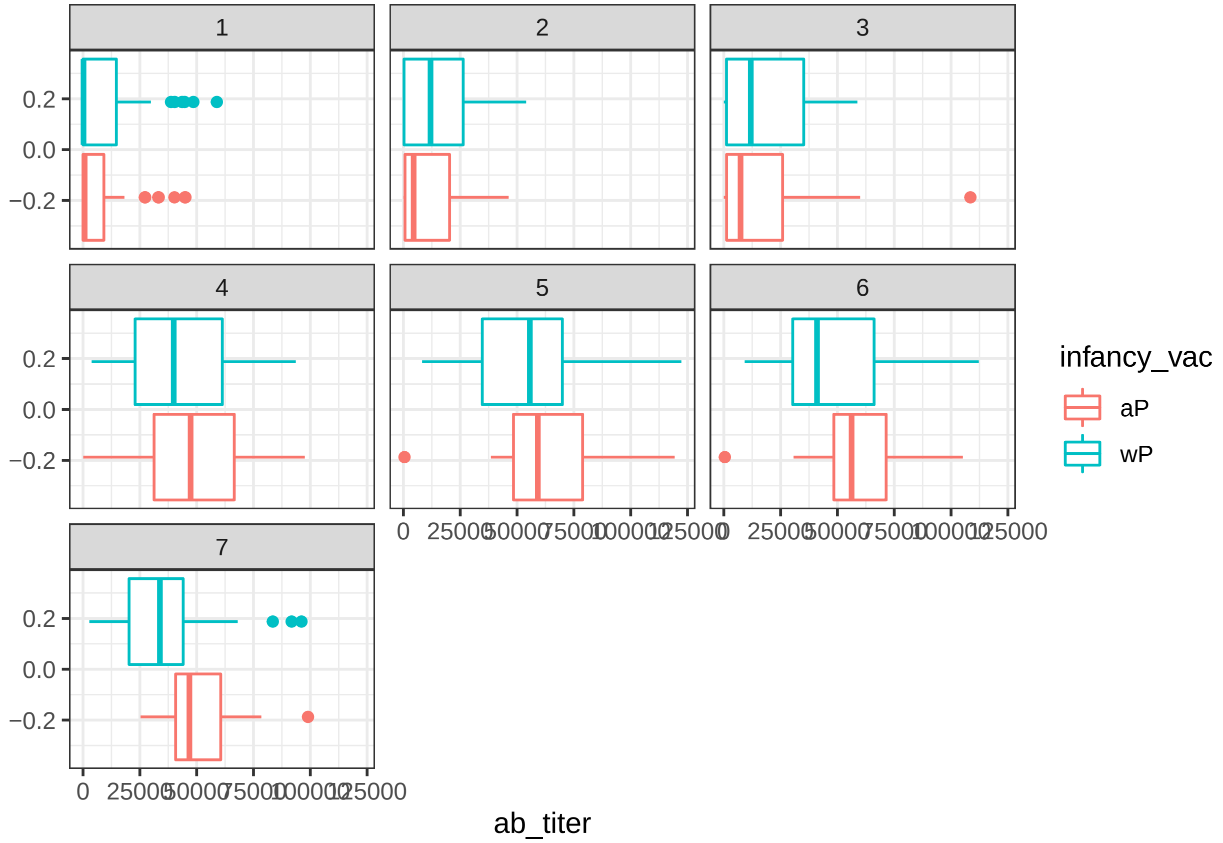
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

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



and antigen=="FIM2/3"

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



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

The measles antigens stay relatively constant over time, and the FIM2/3 antigen rises significantly over time, hitting its peak at day 5 and start decreasing again

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

There is not much of a clear difference between aP and wP responses. Possibly the wP antigens rise and drop quicker

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)
```

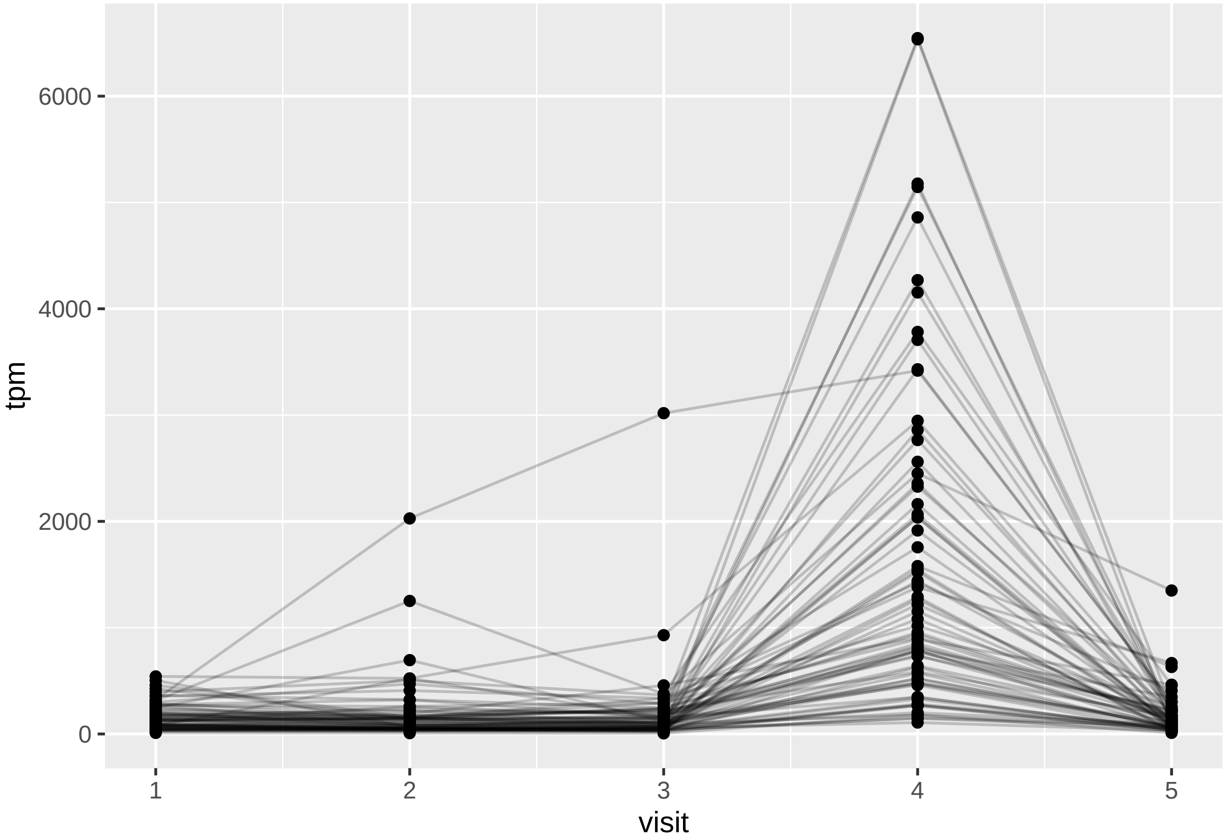
To facilitate further analysis we need to "join" the rna expression data with our metadata meta, which is itself a join of sample and specimen data. This will allow us to look at this genes TPM expression values over aP/wP status and at different visits (i.e. times):

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

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

> 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)?

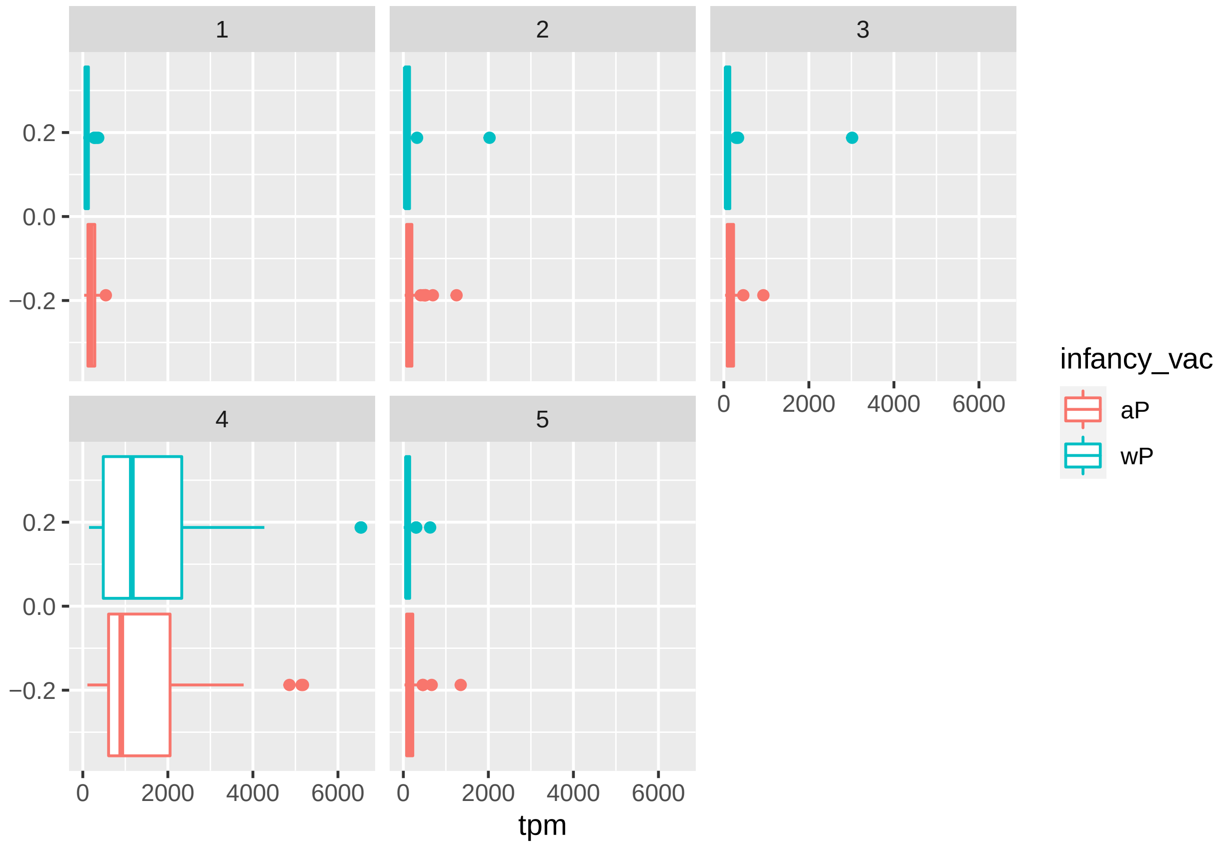
The expression of this gene is at its highest at visit 4, then drops back down at visit 5 to about the same level it was for visits 1-3

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

The antibody titer data shows that antibodies are the highest at visit 5. It makes sense that gene expression is highest at level 4, and by the time of visit 5 the antibodies are made and are the highest then

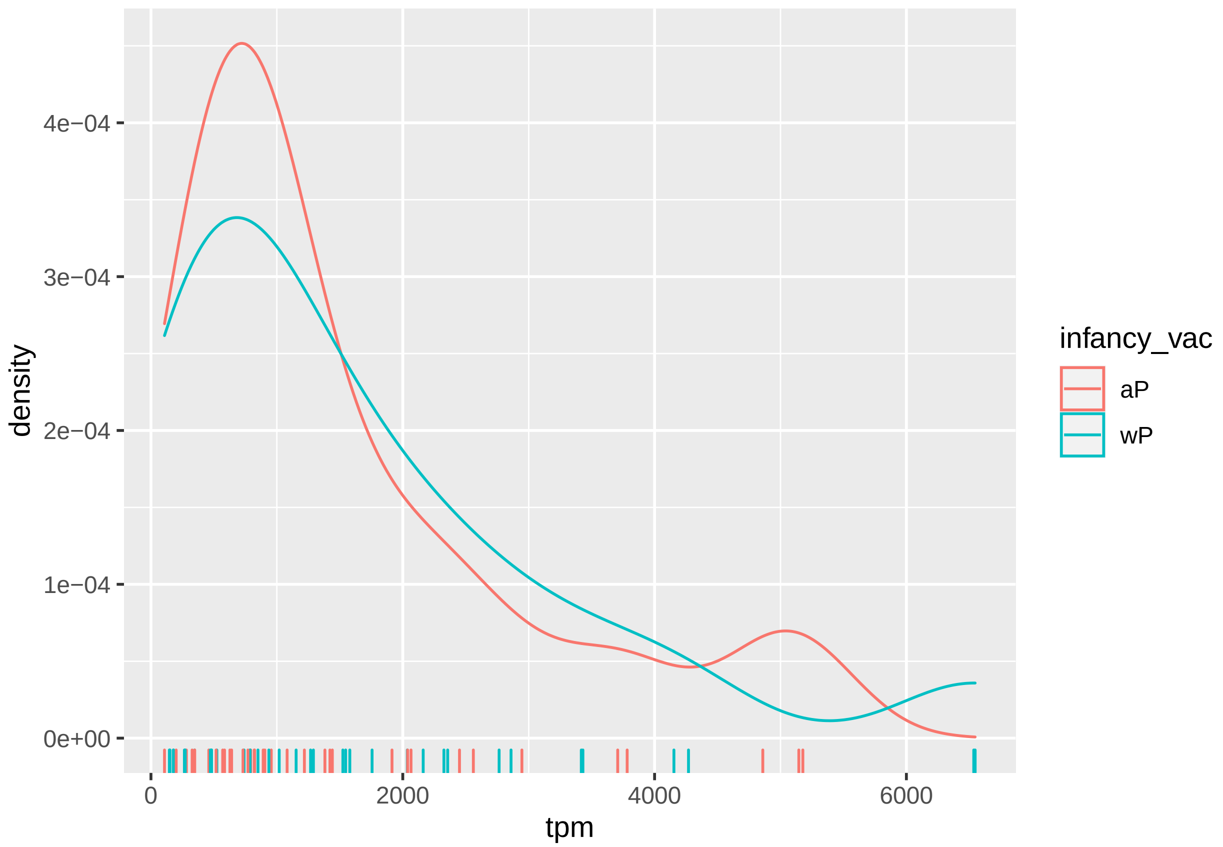
facet by infancy_vac status

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



There is no obvious wP vs. aP differences here even if we focus in on a particular visit:

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



Section 6: Working with larger datasets

We will take their “2020 longitudinal RNA-Seq data” file (named 2020LD_rnaseq.csv) from here: https://www.cmi-pb.org/downloads/cmipb_challenge_datasets/2021_cmipb_challenge/

```
rnaseq <- read.csv("2020LD_rnaseq.csv")
head(rnaseq,3)
```

```
##   versioned_ensembl_gene_id specimen_id raw_count tpm
## 1        ENSG00000229704.1         209         0   0
## 2        ENSG00000229707.1         209         0   0
## 3        ENSG00000229708.1         209         0   0
```

```
dim(rnaseq)
```

```
## [1] 10502460        4
```

how many genes we have reported for each specimen_id

```
n_genes <- table(rnaseq$specimen_id)
head( n_genes , 10)
```

```
## 
##     1     3     4     5     6    19    20    21    22    23 
## 58347 58347 58347 58347 58347 58347 58347 58347 58347 58347
```

how many specimens?

```
length(n_genes)
```

```
## [1] 180
```

Check if there are the same number of genes for each specimen

```
all(n_genes[1]==n_genes)
```

```
## [1] TRUE
```

let's convert to wider format with the pivot_wider() function from the tidyr package:

```
library(tidyr)

rna_wide <- rnaseq %>%
  select(versioned_ensembl_gene_id, specimen_id, tpm) %>%
  pivot_wider(names_from = specimen_id, values_from=tpm)

dim(rna_wide)
```

```
## [1] 58347   181
```

```
head(rna_wide[,1:7], 3)
```

```
## # A tibble: 3 x 7
##   versioned_ensembl_gene_id `209` `74` `160` `81` `102` `163`
##   <chr>                     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 ENSG00000229704.1             0     0     0     0     0     0
## 2 ENSG00000229707.1             0     0     0     0     0     0
## 3 ENSG00000229708.1             0     0     0     0     0     0
```

we can filter this to answer key questions

1. Is RNA-Seq expression levels predictive of Ab titers?
2. What differentiates aP vs wP primed individuals?
3. What about decades after their first immunization?