Class 15 Mini-Project: Investigating Pertussis Resurgence

Divya Shetty (A15390408)

3/8/2022

Investigating Pertussis Cases by Year (US)

The CDC has tracked the case numbers for pertussis since the 1920s! Information can be viewed at this link: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

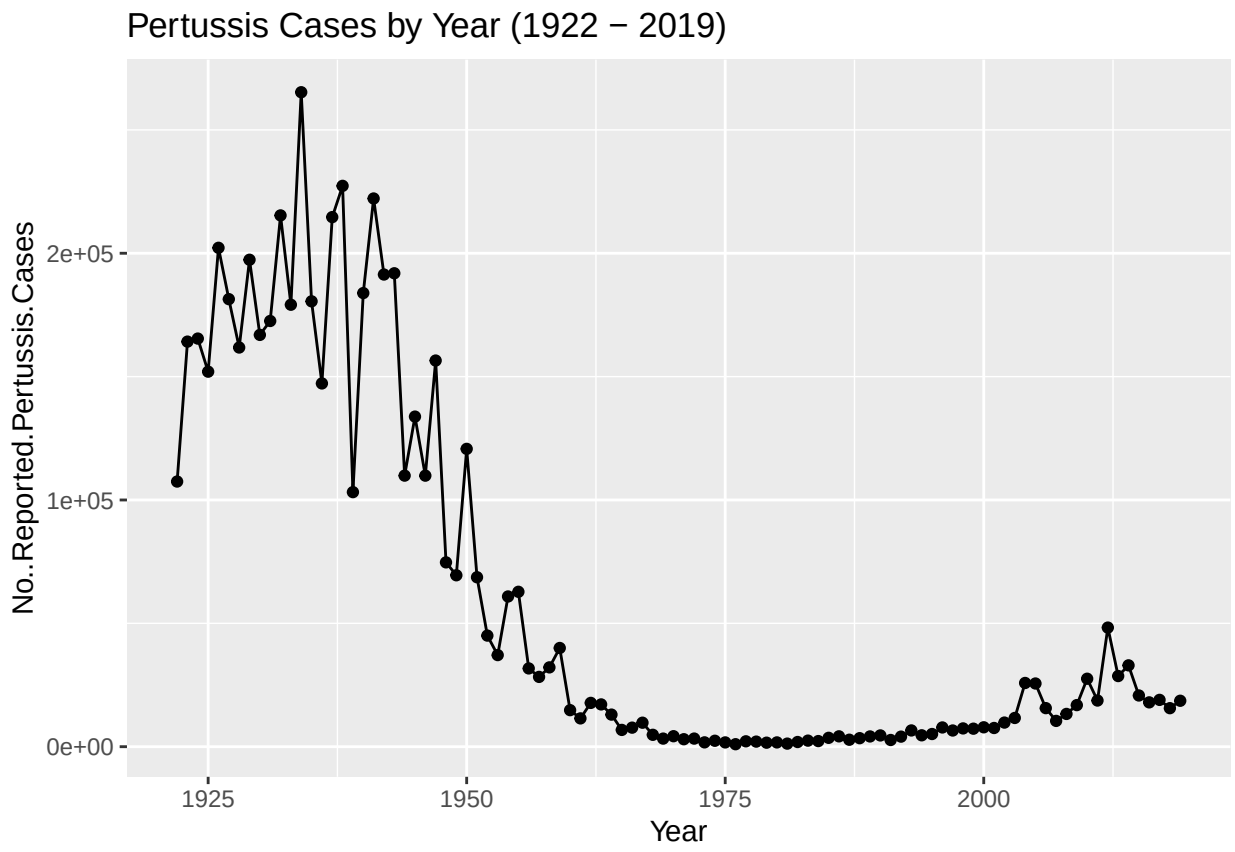
Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
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(x = Year, y = No..Reported.Pertussis.Cases) +
  geom_point() +
  labs(title = "Pertussis Cases by Year (1922 - 2019)",
       xlab = "Time", ylab = "No. of Reported Pertussis Cases") +
  geom_line()
```

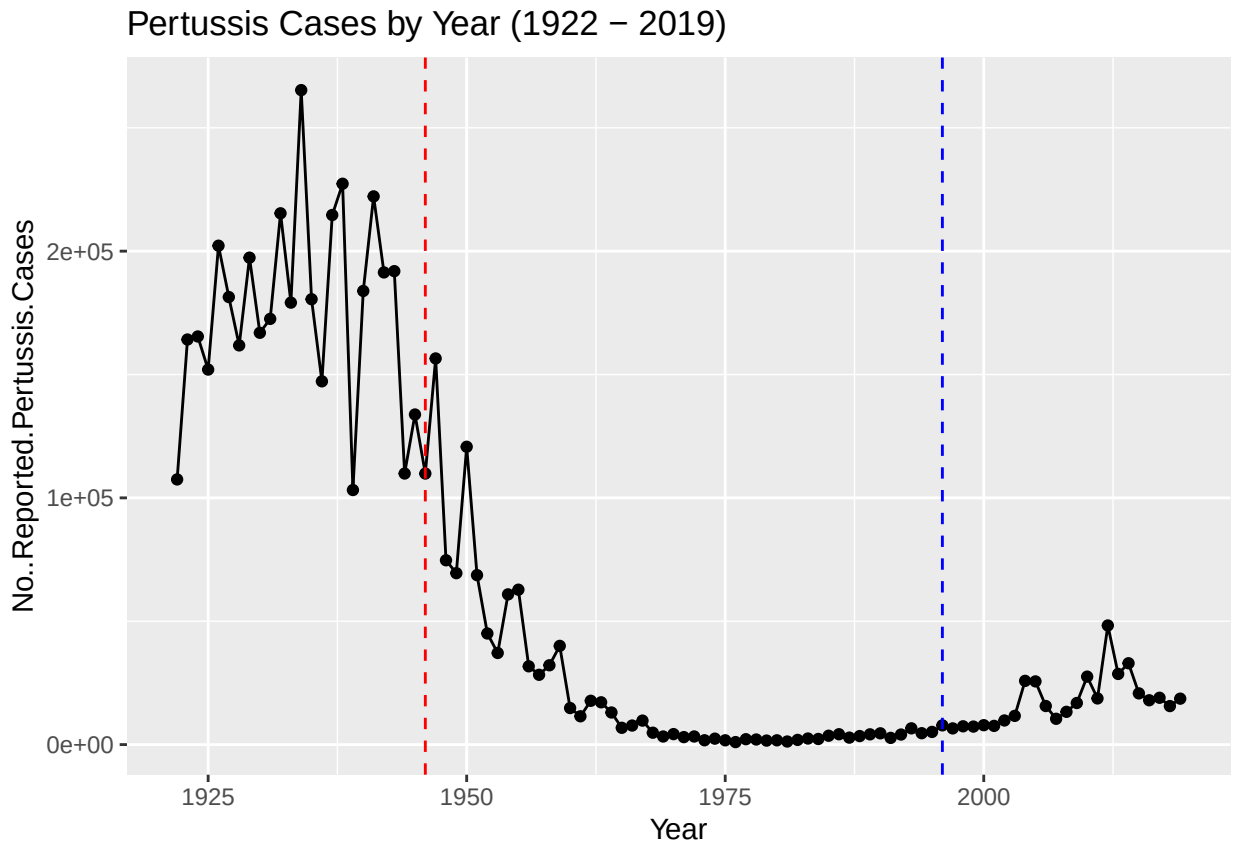


A Tale of Two Vaccine (wP & aP)

Two types of pertussis vaccines are currently available: whole-cell pertussis (wP) and acellular pertussis (aP). The first vaccines were composed of 'whole cell' (wP) inactivated bacteria. The latter aP vaccines use purified antigens of the bacteria and were developed to have less side effects.

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() +
  labs(title = "Pertussis Cases by Year (1922 - 2019)",
       xlab = "Time", ylab = "No. of Reported Pertussis Cases") +
  geom_line() +
  geom_vline(xintercept = 1946, color = "red", linetype = "dashed") +
  geom_vline(xintercept = 1996, color = "blue", linetype = "dashed")
```



After the first vaccine, there appears to be a decrease in cases, implying the vaccine is effective. However, the second vaccine shows a small increase in cases after some time.

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

After the aP vaccine, there was a lag in cases then a sudden spike, This could be due to the development of resistance to the vaccine by the bacteria, which would decrease the effectiveness of the vaccine. Other reasons could include the rise in anti-vaxxers, the increased sensitivity of PCR-based testing, or the waning of immunity in adolescents that had been primed as infants.

Exploring CMI-PB Data

The CMI-PB tracks and makes freely available long-term humoral and cellular immune response data for a large number of individuals who received either DTwP or DTaP combination vaccines in infancy followed by Tdap booster vaccinations. The Tdap vaccine is used as a proxy to measure the immune response for the pertussis vaccines. This data can be used to determine why pertussis, a preventable disease, is on the rise.

CMI-PB API returns JSON data

The API sends data in JSON format, which involves a key-value pair. As such, the jsonlite package must be used in order to read this data! The rjson package may be used for more advanced queries.

Read the "subject" database table from the CMI-PB API using the read_json() function. The first argument is the url where the data is located. The second argument ensures the data is returned in the form of a data frame.

```
library(jsonlite)

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

```
##   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
## 4          4          wP           Male Not Hispanic or Latino Asian
## 5          5          wP           Male Not Hispanic or Latino Asian
## 6          6          wP         Female Not Hispanic or Latino 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
## 4    1988-01-01    2016-08-29 2020_dataset
## 5    1991-01-01    2016-08-29 2020_dataset
## 6    1988-01-01    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 infancy vaccinated subjects and 49 wP infancy vaccinated 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 subjects and 30 male subjects.

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
```

Refer to the table above for the breakdown of demographics.

Working with Dates

The lubridate package can help with examining data related to dates and times. It can convert time differences to years!

```
library(lubridate)

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

```
## [1] 22.18754
```

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?

```
#age of subjects in years
subject$age <- time_length(today() - ymd(subject$year_of_birth), "years")
```

```
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
```

```
wP_vacc <- filter(subject, infancy_vac == "wP")
summary(wP_vacc$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.19   31.19   34.19   35.35   39.19   54.19
```

(i) The average age of wP subjects is 35 years old.

```
aP_vacc <- filter(subject, infancy_vac == "aP")
summary(aP_vacc$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.19   24.19   25.19   24.51   25.19   26.19
```

(ii) The average age of aP subjects is 24 years old.

(iii) Yes, they are significantly different since the maximum age for aP individuals is 26 while the minimum age for wP individuals is 27. Since there is no overlap in values, this indicates a significant difference.

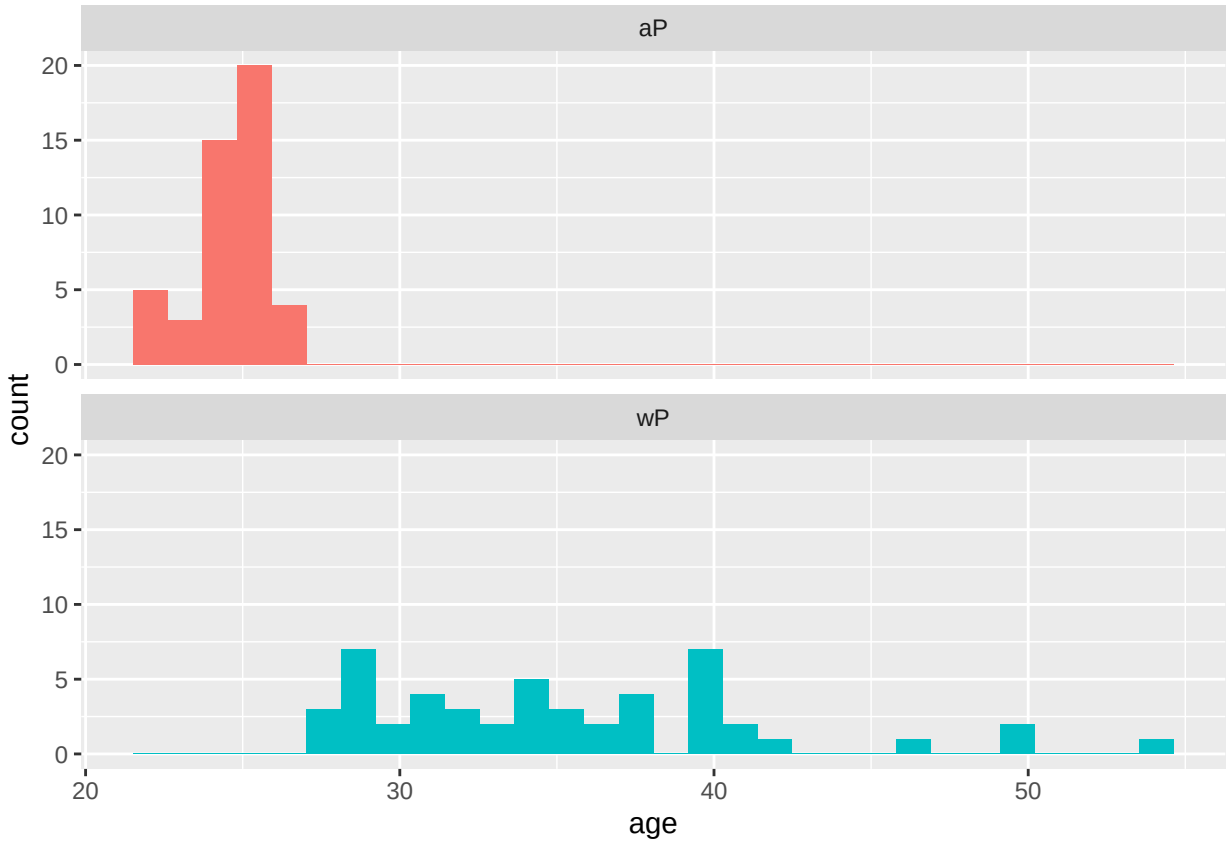
Q8. Determine the age of all individuals at time of boost?

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

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

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

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




```
x <- t.test(wP_vacc$age, aP_vacc$age)

x$p.value
```

```
## [1] 1.316045e-16
```

Yes, I believe the ages are significantly difference as the plots visually demonstrate that there are no overlaps in ages between aP and wP subjects. The calculated p-value is less that 0.05 which indicates a significant difference!

Joining Multiple Tables

```
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)
```

Let's take a quick look at our data!

```
head(specimen)
```

```
##   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
## 1                             0         Blood     1
## 2                           736         Blood    10
## 3                             1         Blood     2
## 4                             3         Blood     3
## 5                             7         Blood     4
## 6                            14         Blood     5
```

To know whether a given specimen_id comes from an aP or wP subject, we need to link 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 36.18617
## 2 36.18617
```

```
## 3 36.18617
## 4 36.18617
## 5 36.18617
## 6 36.18617
```

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
```

```
head(abdata)
```

```
##   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
## 4           1     IgG                TRUE     PRN  332.12718 IU/ML
## 5           1     IgG                TRUE     FHA 1887.12263 IU/ML
## 6           1     IgE                TRUE     ACT    0.10000 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                      NaN          1                           -3
## 2                29.170000          1                           -3
## 3                 0.530000          1                           -3
## 4                 1.070000          1                           -3
## 5                 0.064000          1                           -3
## 6                 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   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 36.18617
## 2 36.18617
## 3 36.18617
## 4 36.18617
## 5 36.18617
## 6 36.18617
```

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
```

Refer to the table above for number of specimens for each isotype.

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
```

The number of specimen for visit 8 is much lower than any other visit. This is because the data is unfinished! It hasn't been collected yet!

Examine IgG1 Ab titer levels

Filter the "abdata" dataset to get the IgG1 isotype, EXCLUDING the values from visit 8 (because the values for this visit are unfinished).

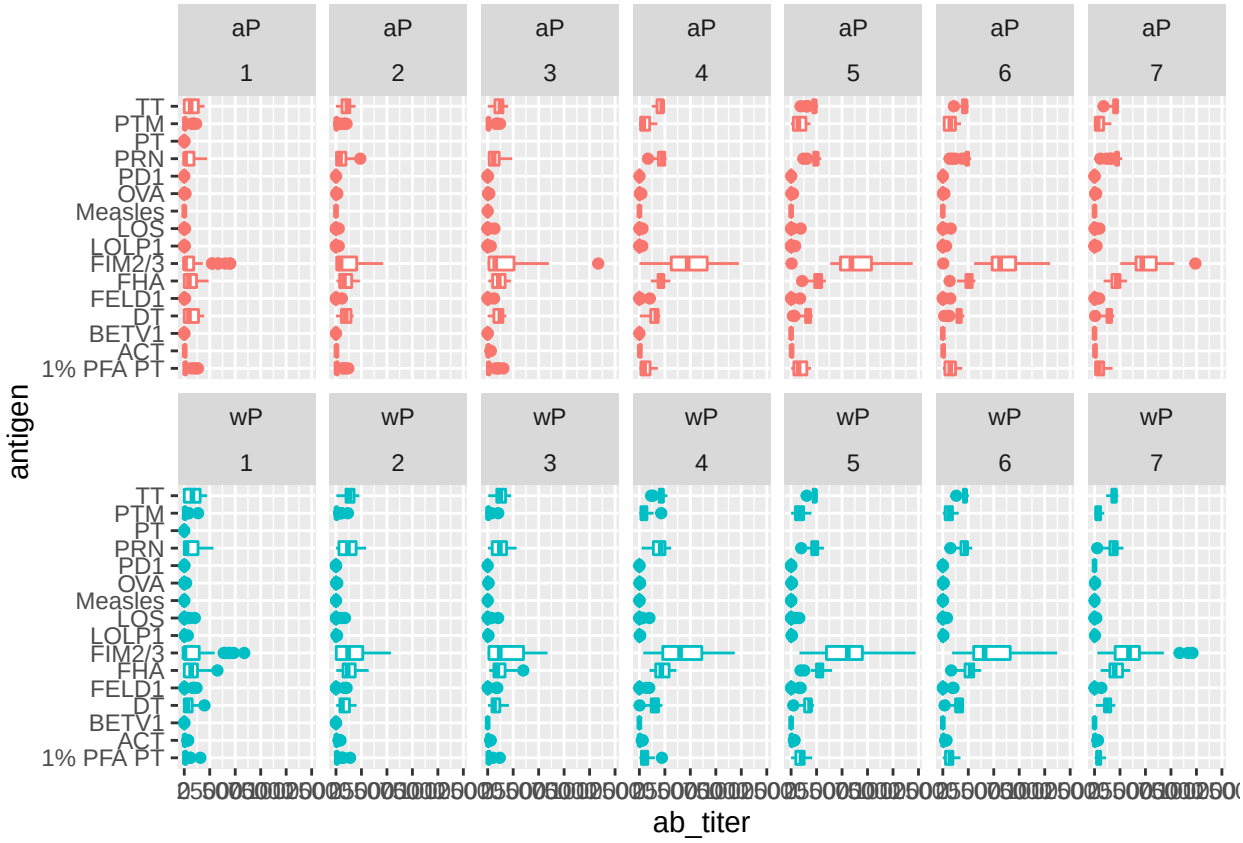
```
ig1 <- filter(abdata, isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   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 36.18617
## 2 36.18617
## 3 36.18617
## 4 36.18617
## 5 36.18617
## 6 36.18617
```

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

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

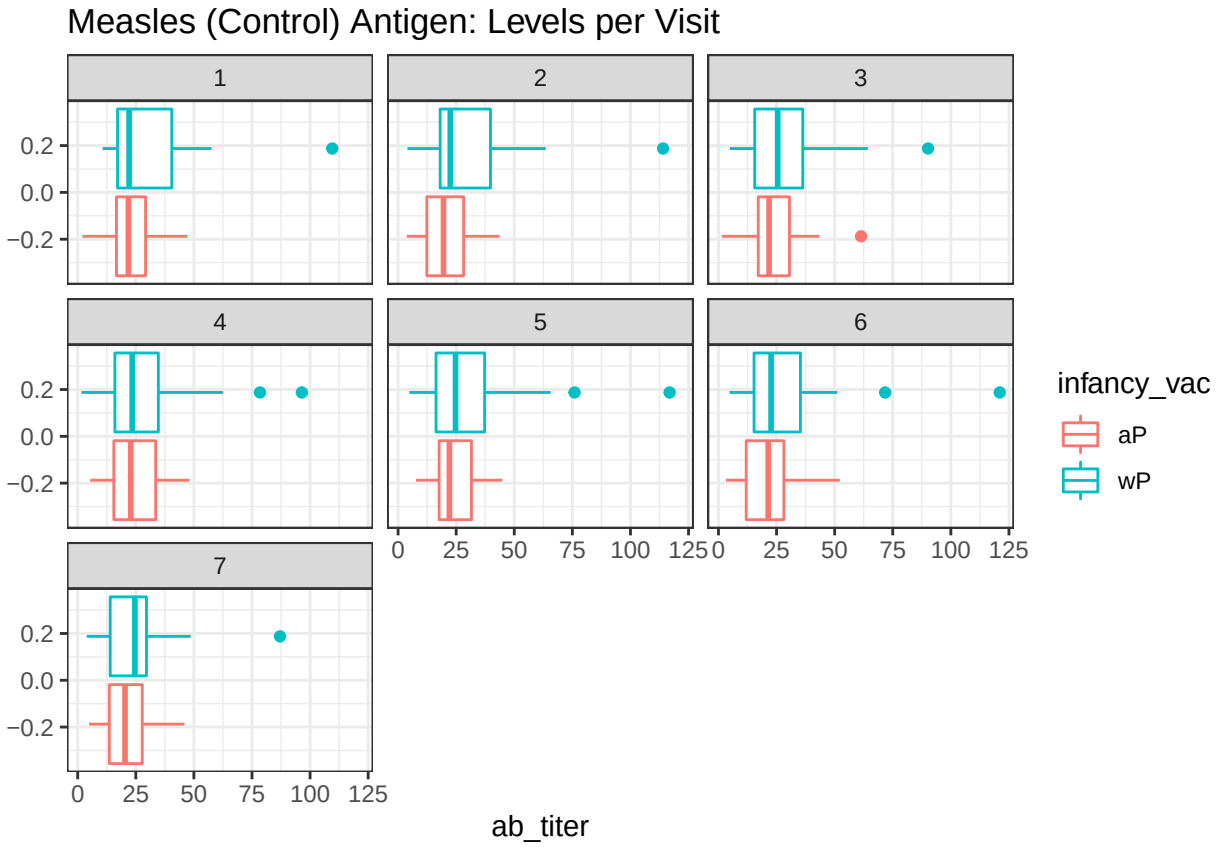


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

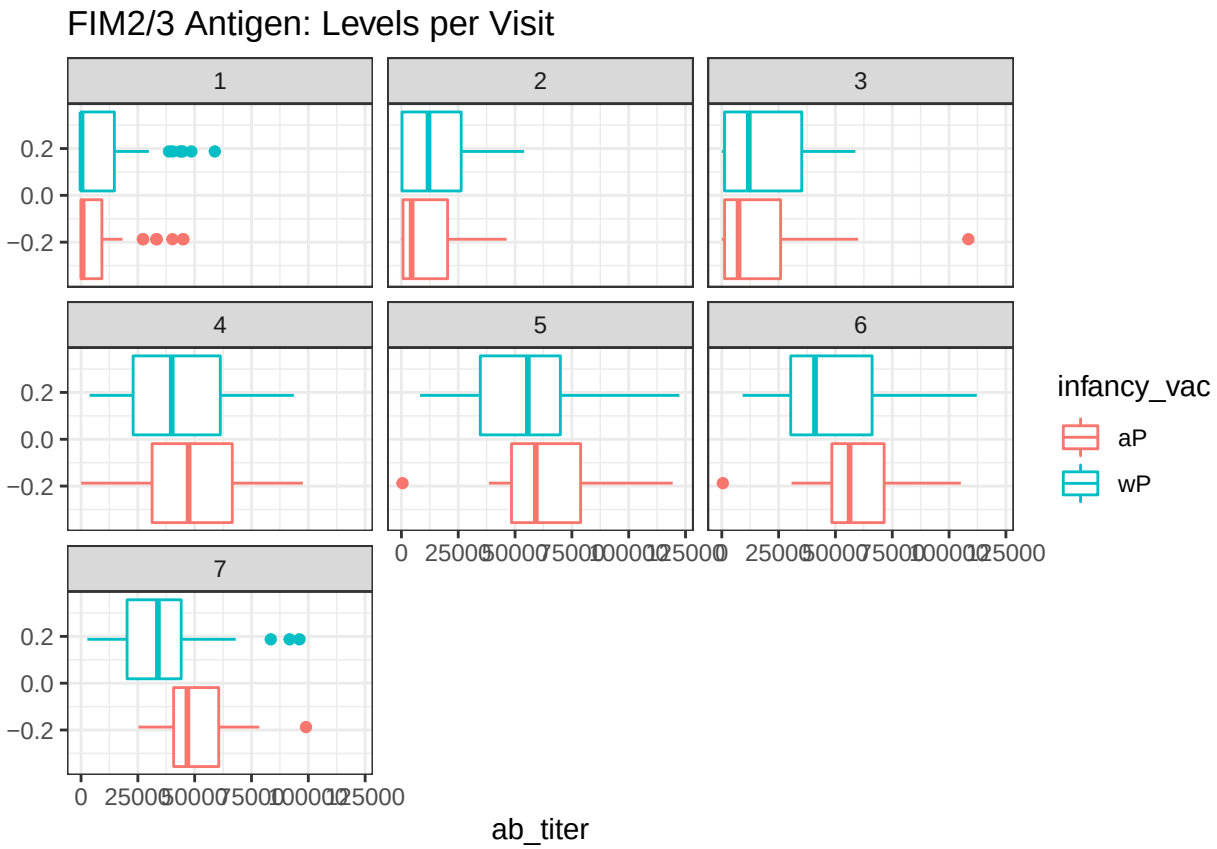
The PTM, PRN, FIM2/3, and FHA antigen appears to show some difference in IgG1 levels, most likely because these antigens are related to components of the bacterium Bordetella pertussis.

Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like.

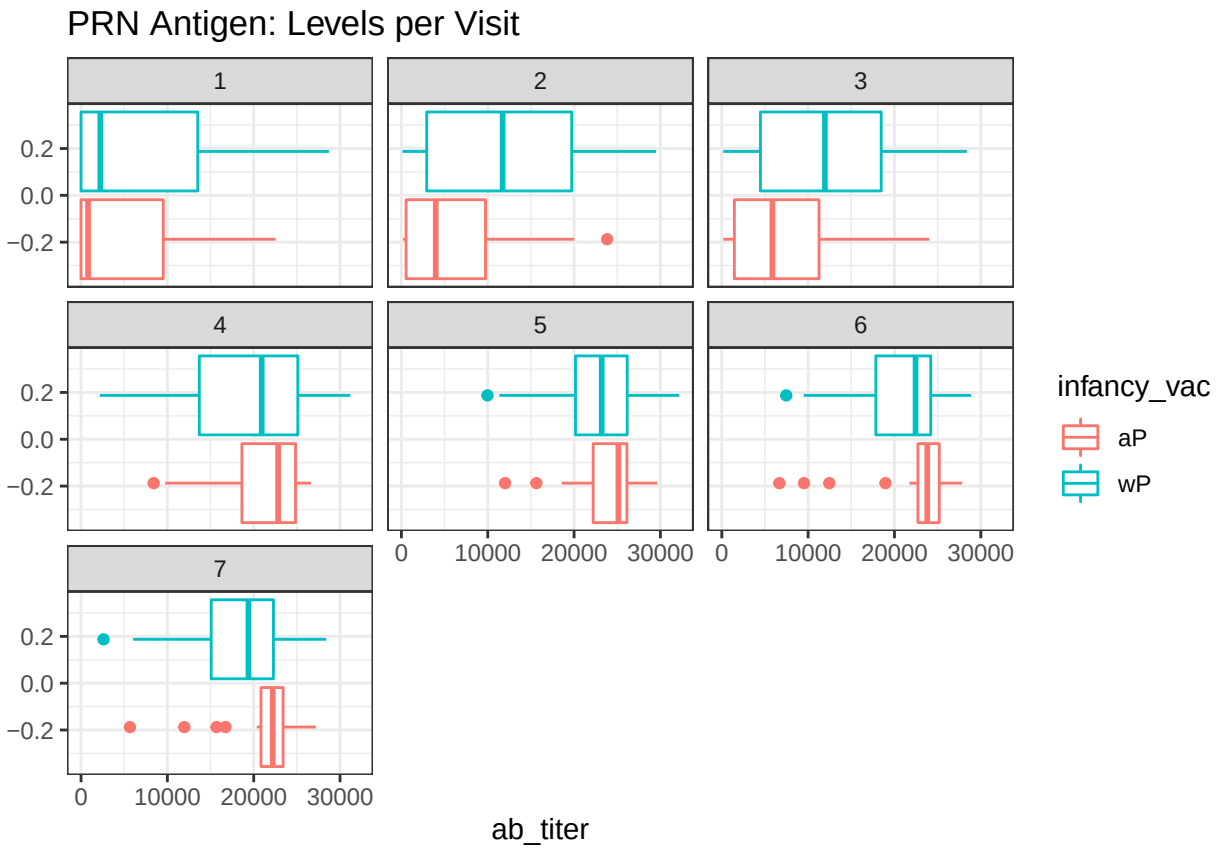
```
filter(ig1, antigen == "Measles") %>%
  ggplot() +
  aes(ab_titer, color = infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "Measles (Control) Antigen: Levels per Visit")
```




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




```
filter(ig1, antigen == "PRN") %>%
  ggplot() +
  aes(ab_titer, color = infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "PRN Antigen: Levels per Visit")
```



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

For both FIM2/3 and PRN, there is a rise in levels over time, whereas the levels for Measles appear to have minimal changes over time. These trends appear to be the same for both aP and wP subjects. The levels for FIM2/3 appear to peak on visit 5.

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

There is no clear difference between aP and wP responses.

Obtaining CMI-PB RNASeq Data

For RNA-Seq data, the API query mechanism will hit the web browser interface limit for file size. HOWEVER, we can still do "targeted" RNA-Seq queries. Below is a link to a key gene (IGHG1) involved in expressing an IgG1 antibody.

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

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

Join the "rna" expression data with our metadata "meta", which will allow us to look at the gene's TPM expression values in relation to aP/wP status and at different visits:

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

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

```
dim(ssrna)
```

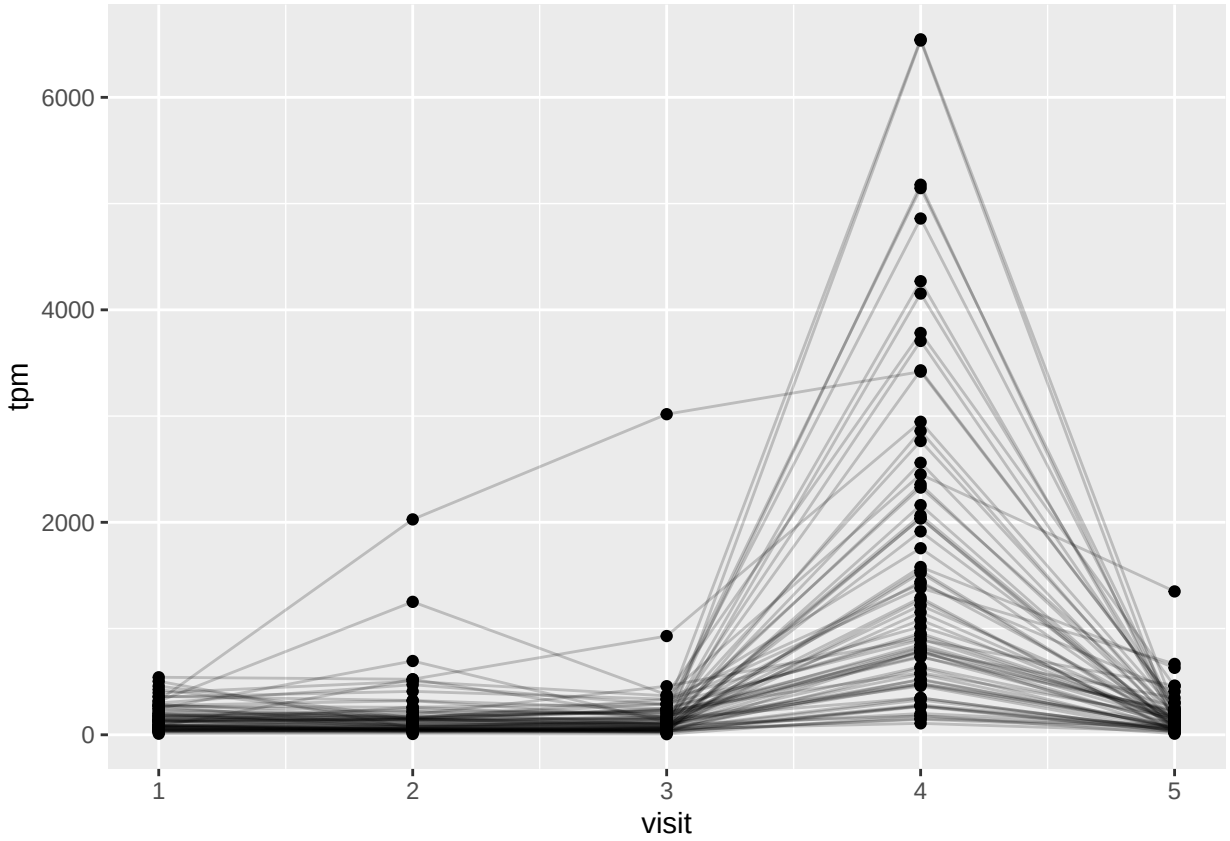
```
## [1] 360  17
```

```
head(ssrna)
```

```
##   versioned_ensembl_gene_id specimen_id raw_count      tpm subject_id
## 1         ENSG00000211896.7         344     18613  929.640         44
## 2         ENSG00000211896.7         243      2011  112.584         31
## 3         ENSG00000211896.7         261      2161  124.759         33
## 4         ENSG00000211896.7         282      2428  138.292         36
## 5         ENSG00000211896.7         345     51963 2946.136         44
## 6         ENSG00000211896.7         244     49652 2356.749         31
##   actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
## 1                            3                             3         Blood
## 2                            3                             3         Blood
## 3                           15                            14         Blood
## 4                            1                             1         Blood
## 5                            7                             7         Blood
## 6                            7                             7         Blood
##   visit infancy_vac biological_sex              ethnicity               race
## 1     3          aP         Female     Hispanic or Latino More Than One Race
## 2     3          wP         Female Not Hispanic or Latino              Asian
## 3     5          wP           Male     Hispanic or Latino More Than One Race
## 4     2          aP         Female     Hispanic or Latino              White
## 5     4          aP         Female     Hispanic or Latino More Than One Race
## 6     4          wP         Female Not Hispanic or Latino              Asian
##   year_of_birth date_of_boost   study_name      age
## 1    1998-01-01    2016-11-07 2020_dataset 24.18617
## 2    1989-01-01    2016-09-26 2020_dataset 33.18549
## 3    1990-01-01    2016-10-10 2020_dataset 32.18617
## 4    1997-01-01    2016-10-24 2020_dataset 25.18549
## 5    1998-01-01    2016-11-07 2020_dataset 24.18617
## 6    1989-01-01    2016-09-26 2020_dataset 33.18549
```

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(x = visit, y = 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)?

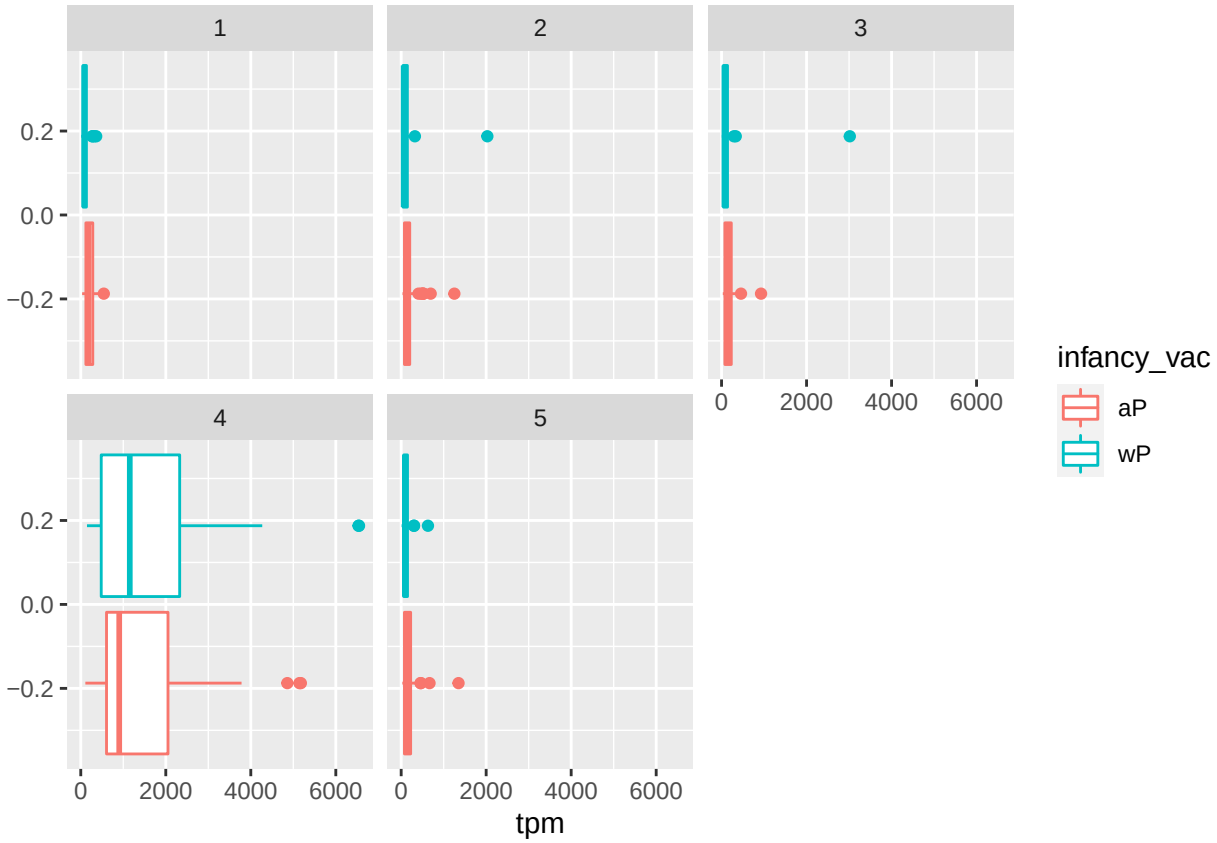
The gene expression appears to peak at visit 4.

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

Yes, it matches what we observed in the antibody data in that the expression levels correspond with a peak in the antibody levels. However, the expression levels drop quickly by visit 5 while the antibody levels remain almost the same. This is because the antibodies created can last.

For fun, facet that data by infancy vaccine status. Is there a difference between aP and wP?

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



Nope! There really isn't a discernible difference!

Working with Larger Datasets [OPTIONAL]

CMI-PB makes CSV files available for download since RNA-Seq queries can be rather large.

```
rnaseq <- read.csv("C:/Users/div/Downloads/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
```

Long Format Data

Note! The "rnaseq" data is in long format, which means genes are in the rows but have we counts for all experiments in one column.

How many genes are 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 are there?

```
length(n_genes)
```

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

Are there the same number of genes for each specimen?

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

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

Convert to Wide Format

```
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
```

Generally, we'd filter out the zero count genese. Since they don't really add to the analysis.

KEY QUESTIONS!!

Is RNA-Seq expression levels predictive of Ab titers?

What differentiates aP vs. wP primed individuals?

What about decades after their first immunization?