Class 19: Investigating Pertussis Resurgence

Angela Liu

1. Investigating Pertussis Cases by Year

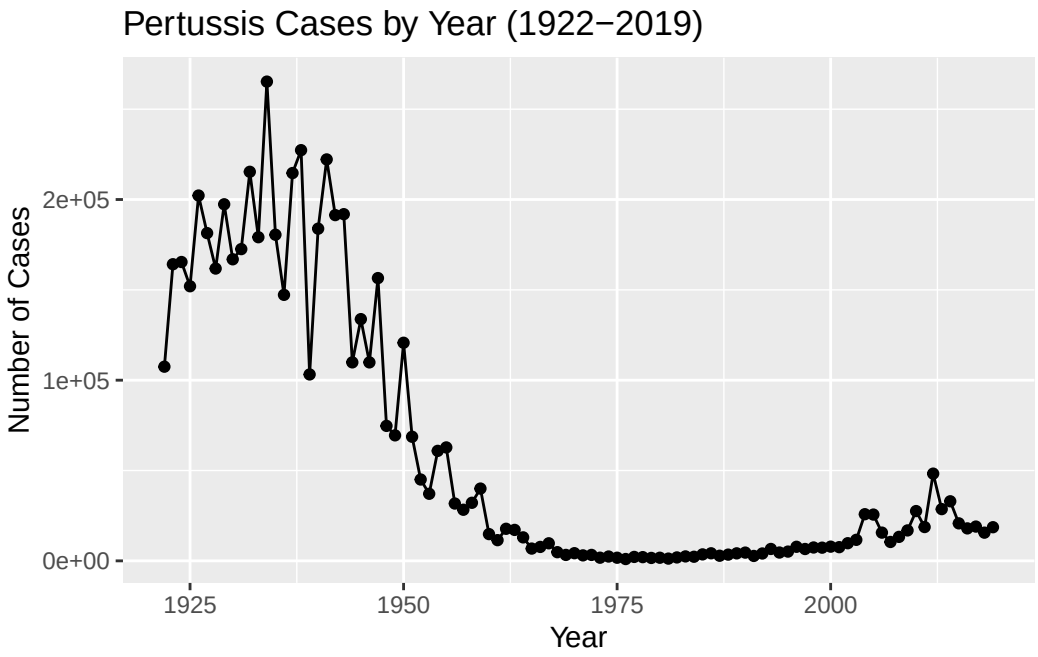
The CDC tracks cases of Pertussis in the US. We can get their data by web-scrapping.

> 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")
```

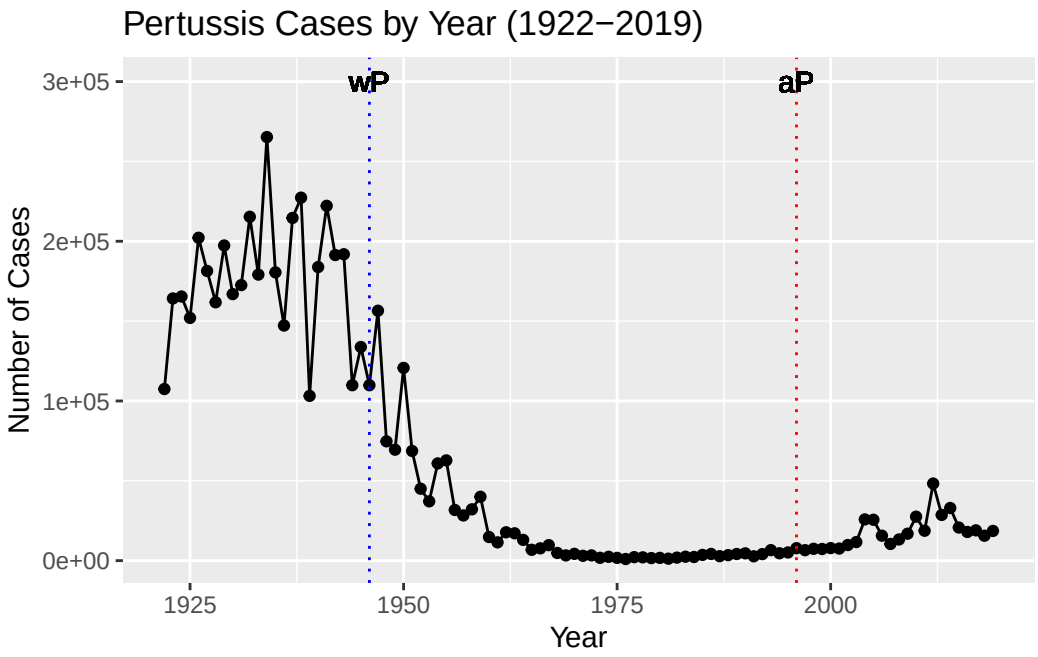
Go to "Addins" in RStudio and paste the CDC data as a data frame.

```
library(ggplot2)
baseplot <- ggplot(cdc, aes(year, cases)) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Number of Cases", title = "Pertussis Cases by Year (1922-2019)")
baseplot
```



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?

```
baseplot + geom_vline(xintercept = 1946, color = "blue", linetype = 3) +
  geom_text(mapping = aes(x = 1946, y = 3e+05, label = "wP")) +
  geom_vline(xintercept = 1996, color = "red", linetype = 3) +
  geom_text(mapping = aes(x = 1996, y = 3e+05, label = "aP"))
```



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, the cases of pertussis remained low but started to increase by a bit. This could be due to the fact that the vaccine rollout took awhile with some more hesitancy as the vaccine is administered at a very young age.

Exploring CMI-PB Data

The CMI-PB project collects data on aP and wP individuals. and their immune response to infection and/or booster shots.

Since the CMI-PB API returns JSON data, let's install the package **jsonlite**.

```
# install.packages("jsonlite")
library(jsonlite)
```

Now we can simplify the JSON key-value pair arrays into R data frames more efficiently.

```
# simplifyVector will return the dataframe as a vector
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 may aP and wP infancy vaccinated subjects are in the dataset?

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

```
aP wP
47 49
```

There are 47 aP 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 females and 30 males.

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
```

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?

```
today()
```

```
[1] "2023-03-14"
```

```
# get the ages of the individuals with today's date
subject$age <- today() - ymd(subject$year_of_birth)
head(subject$age)
```

```
Time differences in days
[1] 13586 20161 14682 12856 11760 12856
```

```
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.age <- filter(subject, infancy_vac == "aP")$age
wp.age <- filter(subject, infancy_vac == "wP")$age

time_length(mean(ap.age), "years")
```

```
[1] 25.5156
```

```
time_length(mean(wp.age), "years")
```

```
[1] 36.36006
```

```
t.test(ap.age, wp.age)
```

```

    Welch Two Sample t-test

data:  ap.age and wp.age
t = -12.092 days, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4618.534 days -3303.337 days
sample estimates:
Time differences in days
mean of x mean of y
 9319.574 13280.510
```

The average age of wP individuals is 36.36 years, of aP individuals is 25.52. The average ages are significantly different.

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

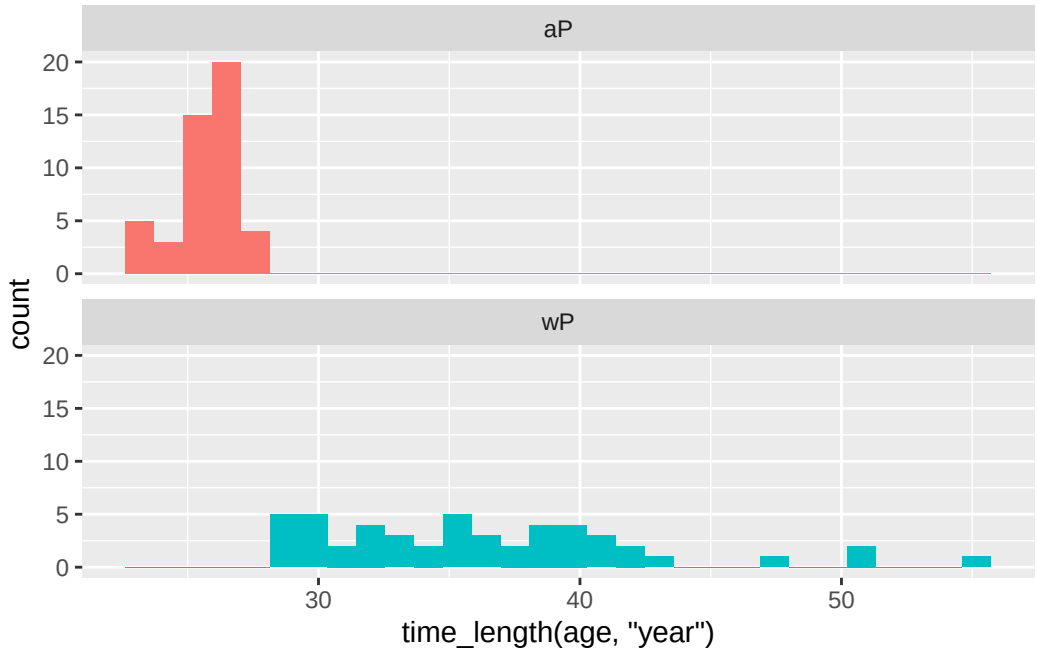
```
boost_time <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(boost_time, "year")
head(age_at_boost)
```

```
[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(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`.
```



Yes, looking at the boxplot, the groups are significantly different as they have very different spreads and centers.

Joining Multiple Tables

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)
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
```

```
head(titer)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
4           1     IgG                TRUE     PRN  332.12718       2.602350
5           1     IgG                TRUE     FHA 1887.12263      34.050956
6           1     IgE                TRUE     ACT    0.10000       1.000000
   unit lower_limit_of_detection
1 UG/ML                2.096133
2 IU/ML               29.170000
3 IU/ML                0.530000
4 IU/ML                6.205949
5 IU/ML                4.679535
6 IU/ML                2.816431
```

To know whether a given specimen_id comes from an aP or wP individual we need to merge our specimen and subject data frames. The **dplyr** package has a family of join() functions that can help us with this common task:

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 with `by = join_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      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 13586 days
2 13586 days
3 13586 days
```

```
4 13586 days
5 13586 days
6 13586 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 with `by = join_by(specimen_id)`
```

```
dim(abdata)
```

```
[1] 32675    21
```

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 six specimens – IgE, IgG, 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 are significantly less number of visit 8 specimens compared to the others.

It is still quite small as it is currently occurring. It is recommended that we remove visit 8 from our analysis.

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.

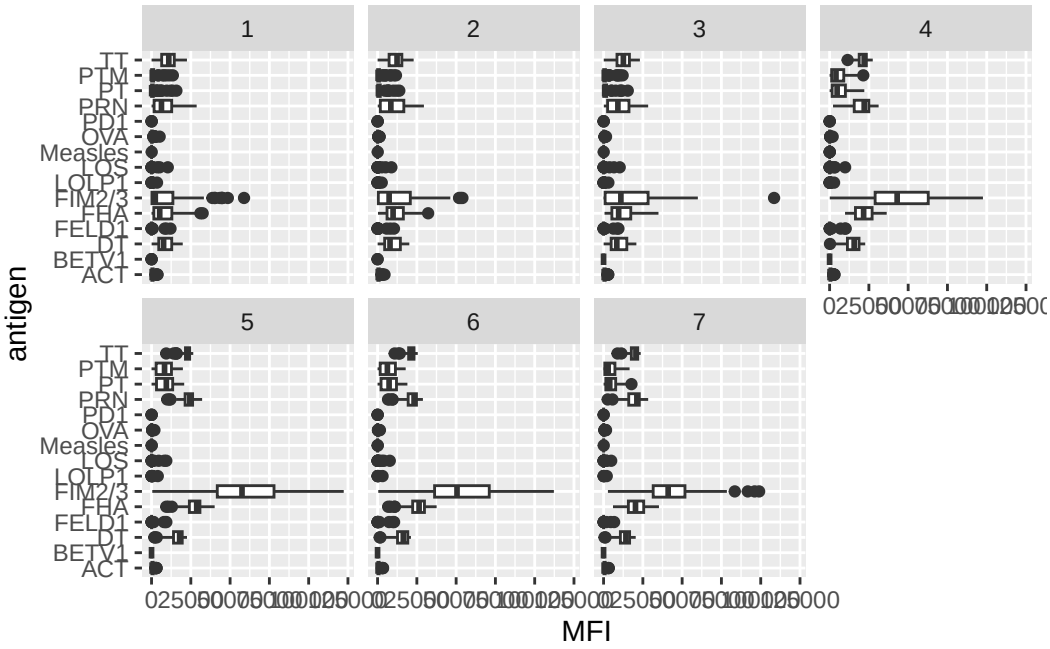
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1    IgG1                TRUE     ACT 274.355068      0.6928058
2           1    IgG1                TRUE     LOS  10.974026      2.1645083
3           1    IgG1                TRUE   FELD1   1.448796      0.8080941
4           1    IgG1                TRUE   BETV1   0.100000      1.0000000
5           1    IgG1                TRUE   LOLP1   0.100000      1.0000000
6           1    IgG1                TRUE Measles  36.277417      1.6638332
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 3.848750          1                           -3
2 IU/ML                 4.357917          1                           -3
3 IU/ML                 2.699944          1                           -3
4 IU/ML                 1.734784          1                           -3
5 IU/ML                 2.550606          1                           -3
6 IU/ML                 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      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 13586 days
2 13586 days
3 13586 days
4 13586 days
5 13586 days
```

```
6 13586 days
```

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

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

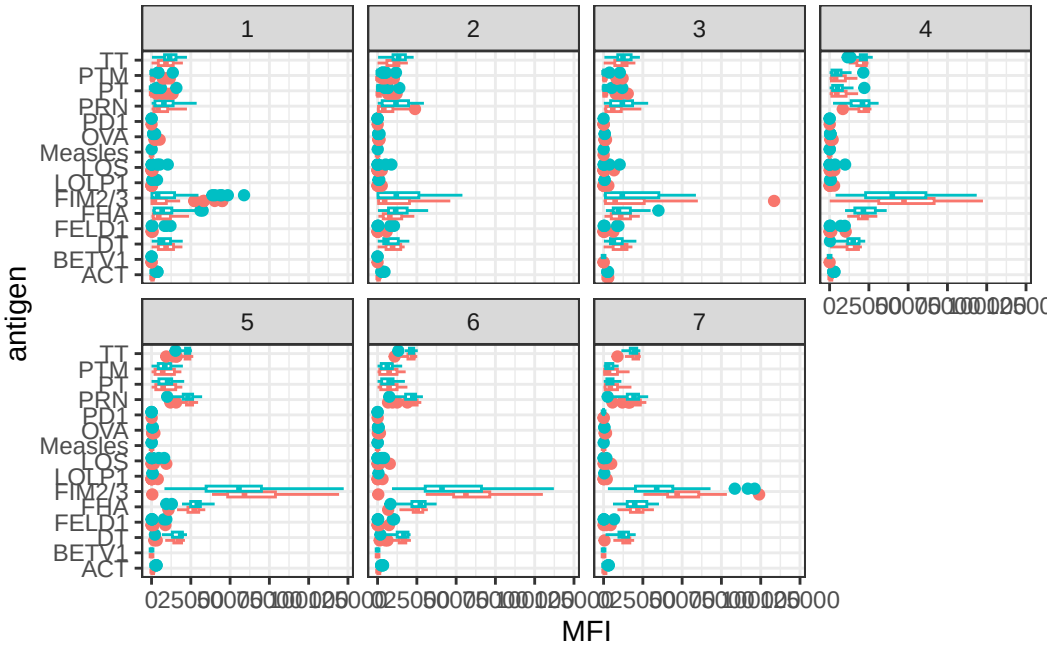


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3 shows the most difference in the level of IgG1 antibody as they are the fibers with the cell adhesion markers. TT (tetanus toxin) also shows some difference in IgG1 antibody as it inhibits the release of neurotransmitters. FHA (filamentous hemagglutinin) plays a role in host-cell binding and infection.

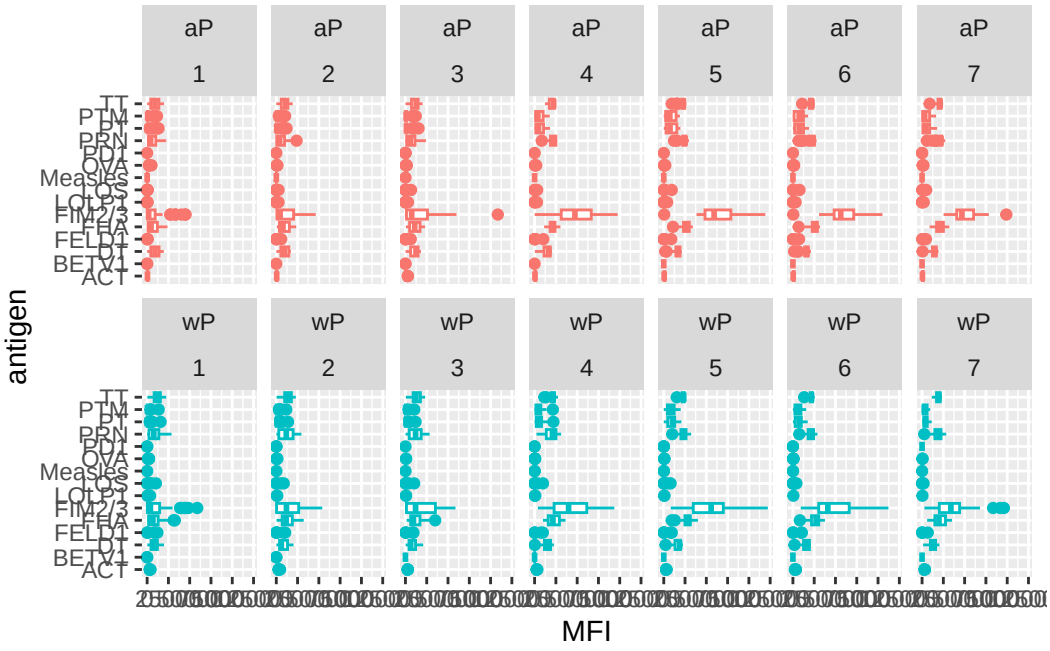
We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status (see below). However these plots tend to be rather busy and thus hard to interpret easily.

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



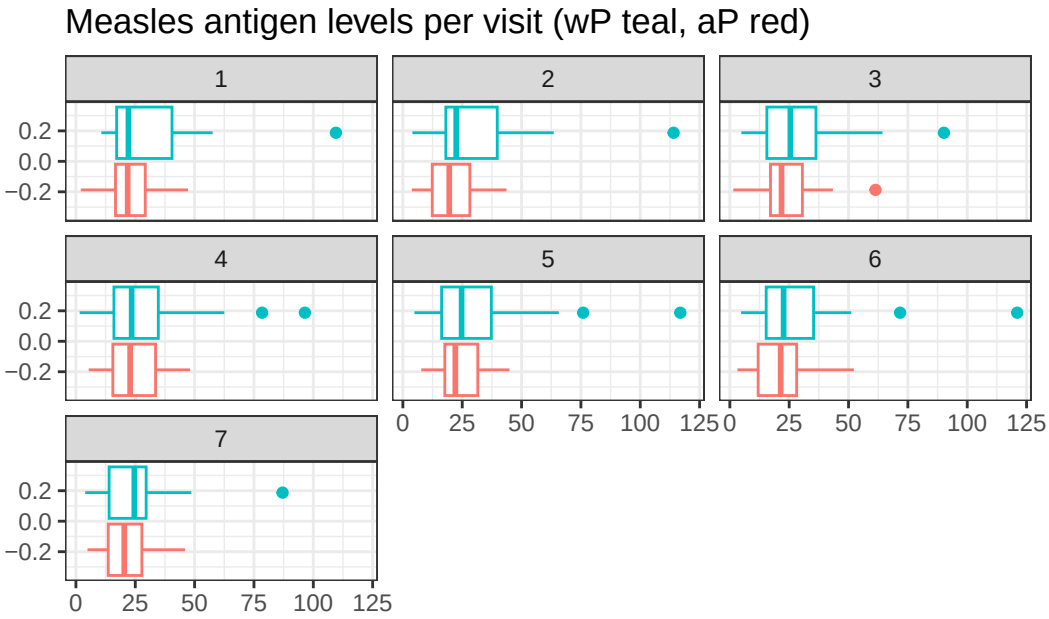
We can add `infancy_vac` to the faceting:

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

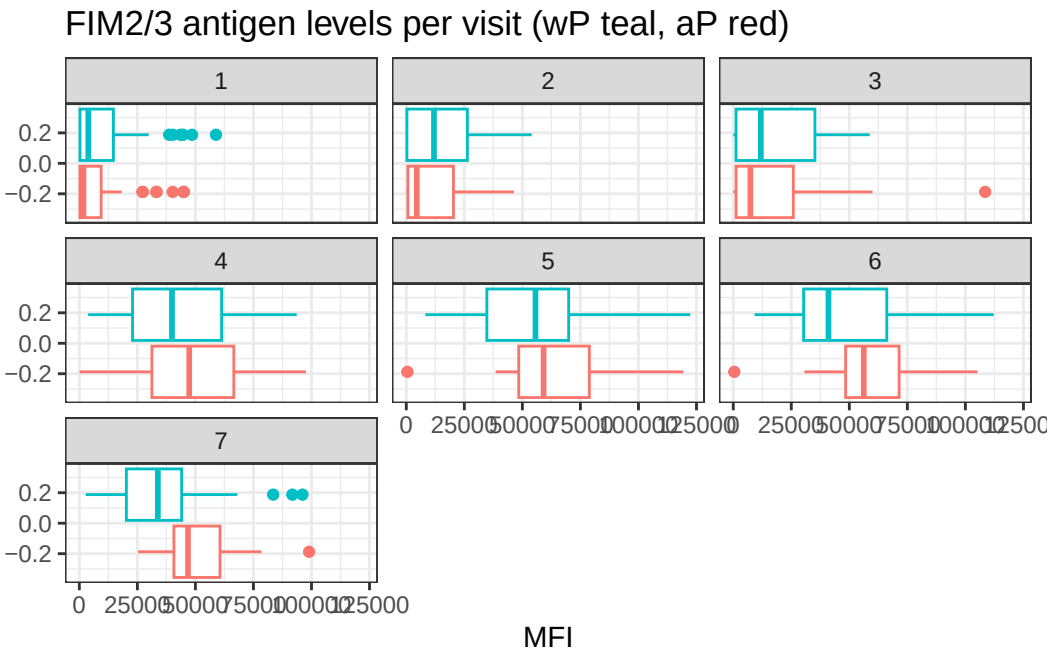


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate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

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  ggtitle("Measles antigen levels per visit (wP teal, aP red)")
```




```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  ggtitle("FIM2/3 antigen levels per visit (wP teal, aP red)")
```



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

FIM2/3 for both aP and wP groups increase over time compared to measles. It seems that FIM2/3 antigen levels reaches a peak at visit 5.

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

There is not too big of a difference. However, aP antigen levels do tend to increase and surpass wP at days 4, 6, 7.

Obtaining CMI-PB RNASeq Data

For RNA-Seq data, the API query mechanism quickly hits the web browser interface limit for file size. We will present alternative download mechanisms for larger CMI-PB datasets in the next section. However, we can still do "targeted" RNA-Seq querys via the web accessible API.

For example we can obtain RNA-Seq results for a specific ENSEMBLE gene identifier or multiple identifiers combined with the & character:

```
# For example use the following URL
# https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7
```

This link is related to the key gene involved in the expression of any IgG1 antibody – specifically the IGHG1 gene.

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

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

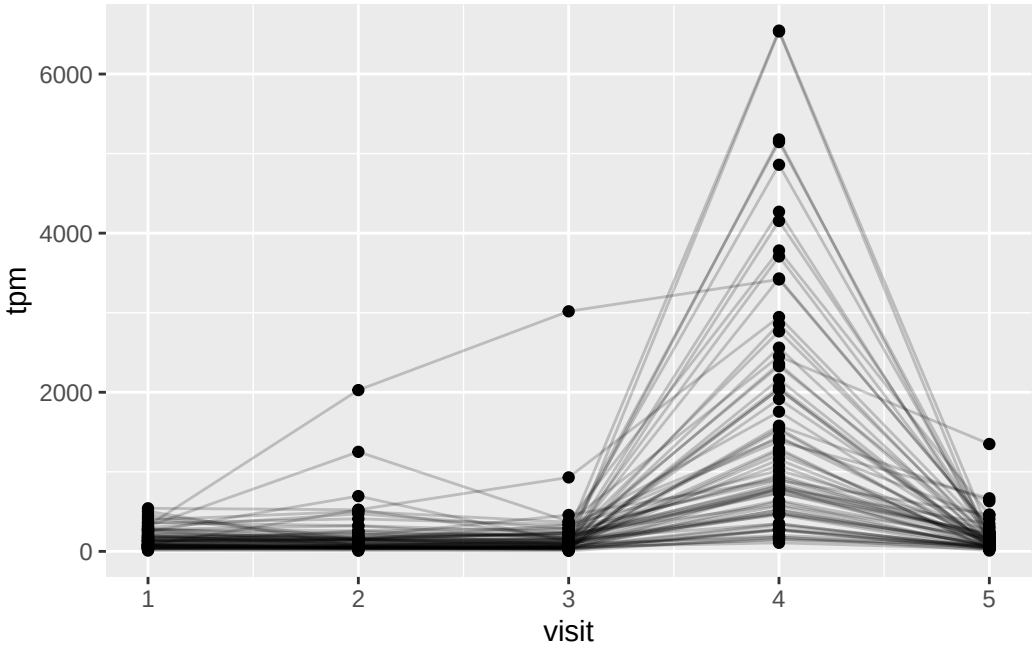
We will join together the `rna` expression data with our metadata of sample and specimen data, `meta`. Now we can look at the genes' TPM expression values over aP/wP status and at different time visits.

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

```
Joining with `by = join_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)?

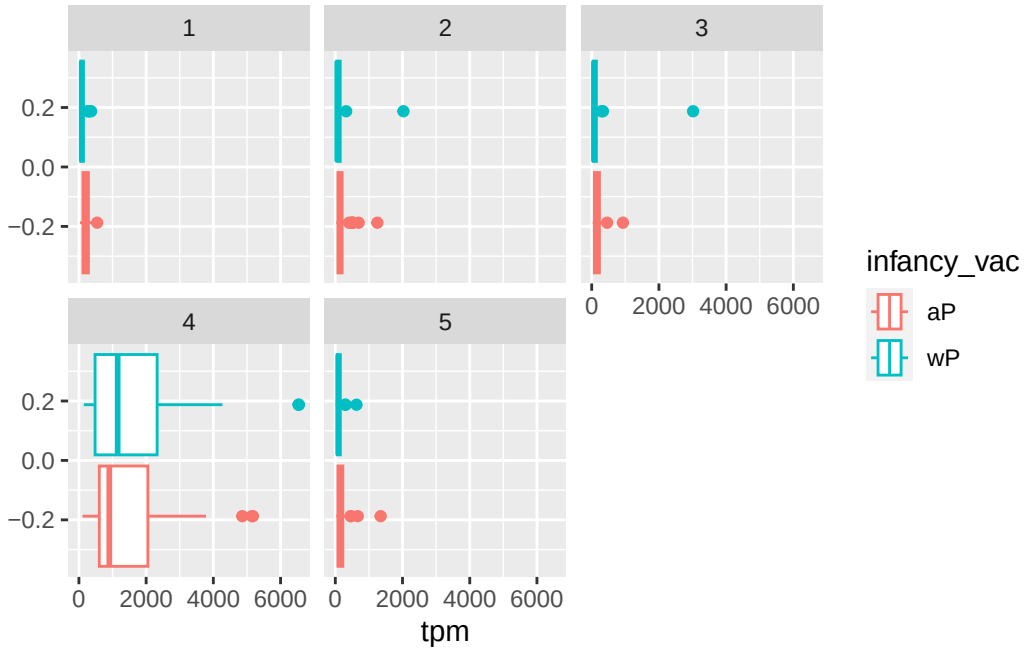
The maximum expression of the gene occurs consistently on visist 4 for the samples. There's a very sharp increase in gene expression from visit 3 to visit 4 and a sharp decline between visit 4 and 5.

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

The patterns do not match. The gene is highly expressed at a certain point whie the cells are making antibodies, which are long lived. THe antibody titer data has a more gradual change as opposed to the sharp peaks of the gene expression.

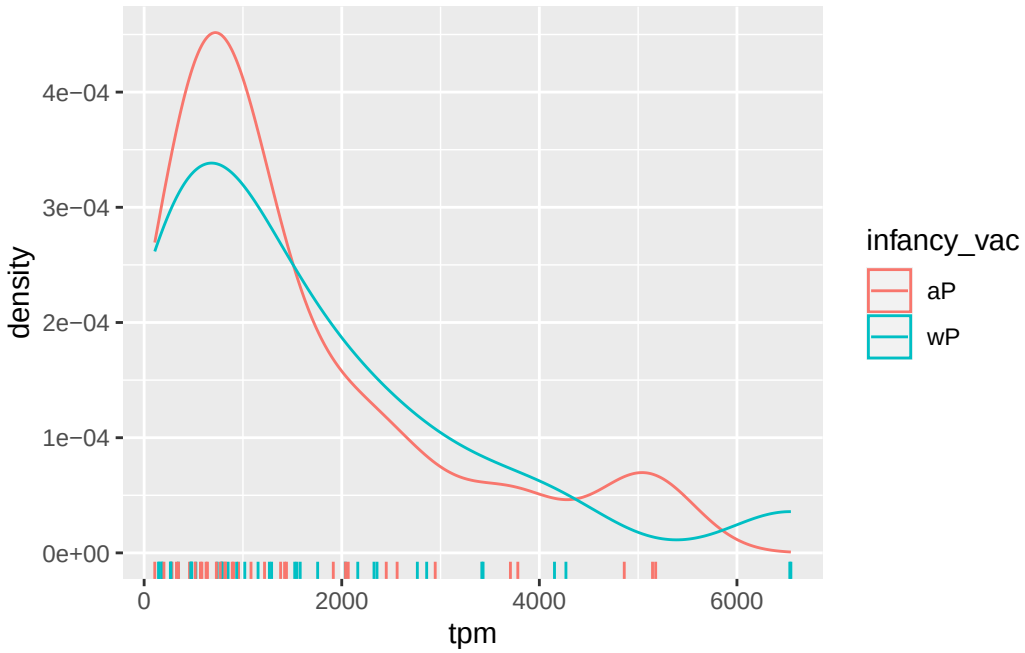
We can dig deeper and color and/or facet by `infancy_vac` status:

```
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()
```


density
4e−04
3e−04
2e−04
1e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP