Class 17: Investigating Pertussis

AUTHOR

Hyeseung (Frankie) Son

Investigating Pertussis Cases by Year

Pertussis, or whooping cough, is a highly contagious lung infection caused by a bacteria *B. pertussis*.

The CDC tracks reported cases in the U.C since the 1920s.

> 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.

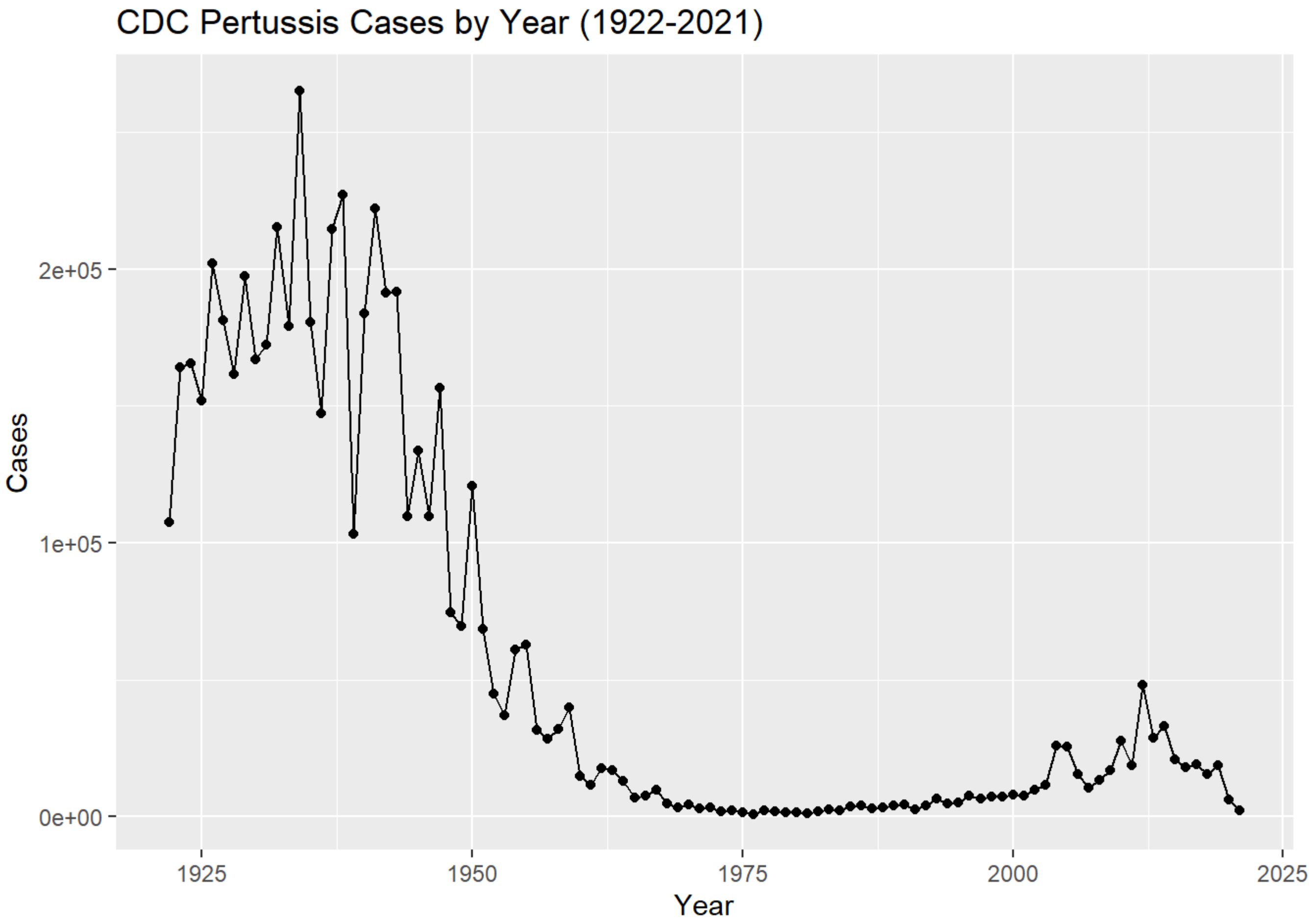
```
cdc <- data.frame(
                                 Year = c(1922L,1923L,1924L,1925L,
                                          1926L,1927L,1928L,1929L,1930L,1931L,
                                          1932L,1933L,1934L,1935L,1936L,
                                          1937L,1938L,1939L,1940L,1941L,1942L,
                                          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, 2020L, 2021L),
         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,
                                           20762,17972,18975,15609,18617,6124,2116)
       )
```

We can now plot the number of reported pertussis cases per year in the U.S.

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
ggplot(cdc) +
  aes(x = Year, y = Cases ) +
  ggtitle("CDC Pertussis Cases by Year (1922-2021)") +
  geom_point() +
  geom_line()
```



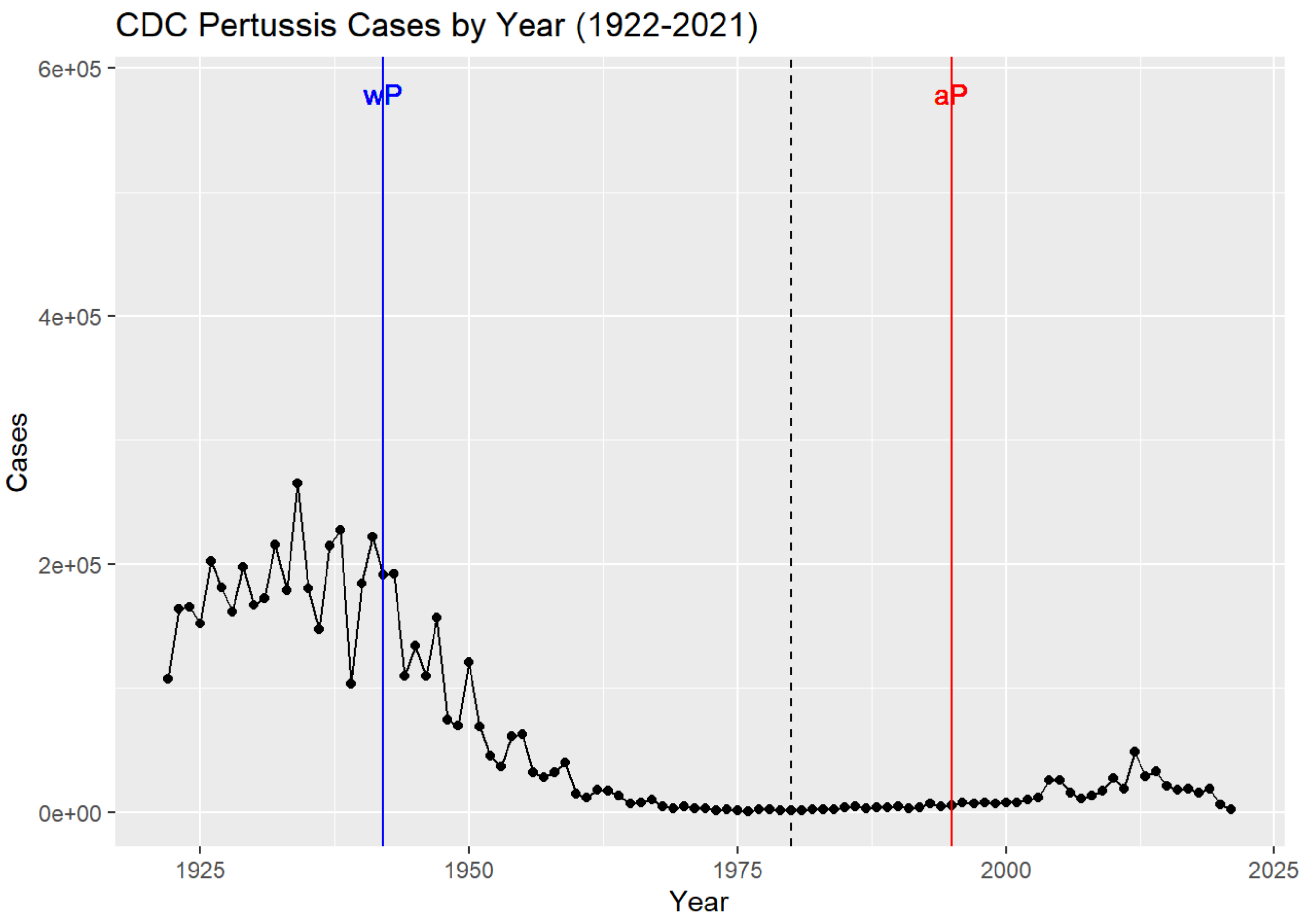
The first big whole-cell pertussis vaccine program started in 1942.

Examining 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?

The 1948 introduction of the wP vaccine leads to a significant decrease in cases for a couple decades. There is another increase in pertussis cases after introduction of the aP vaccine.

```
library(ggplot2)

ggplot(cdc) +
  aes(x = Year, y = Cases ) +
  ggtitle("CDC Pertussis Cases by Year (1922-2021)") +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1942, color = "blue") +
  geom_vline(xintercept = 1980, color = "gray3", linetype = 2) +
  geom_vline(xintercept = 1995, color = "red" ) +
  geom_text(data = cdc, aes(x = 1942, y = 580000, label = "wP"), color = "blue") +
  geom_text(data = cdc, aes(x = 1995, y = 580000, label = "aP"), color = "red")
```



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

The bacteria may have gone through evolution by uptake of foreign DNA and developed a resistance to the vaccine, waning vaccine efficacy, which explains the increase in reported cases.

Something big is happening with pertussis cases and big outbreaks are once again a major public health concern!

Exploring CMI-PB data

Enter the CMI-PB Project, which is studying this problem on large scale. Let’s see what data they have.

Their data is available in JSON format (“key:value” pair style). We will use “jsonlite” package to read their data.

```r
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.2.3
```

```r
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      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
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?

```r
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
```

> 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?

```
library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.2.3
```

```
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.2.3
```

```
subject$age <- today() - ymd(subject$year_of_birth)

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     23      25      26      26      26      27 
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     28      32      35      37      40      55 
```

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

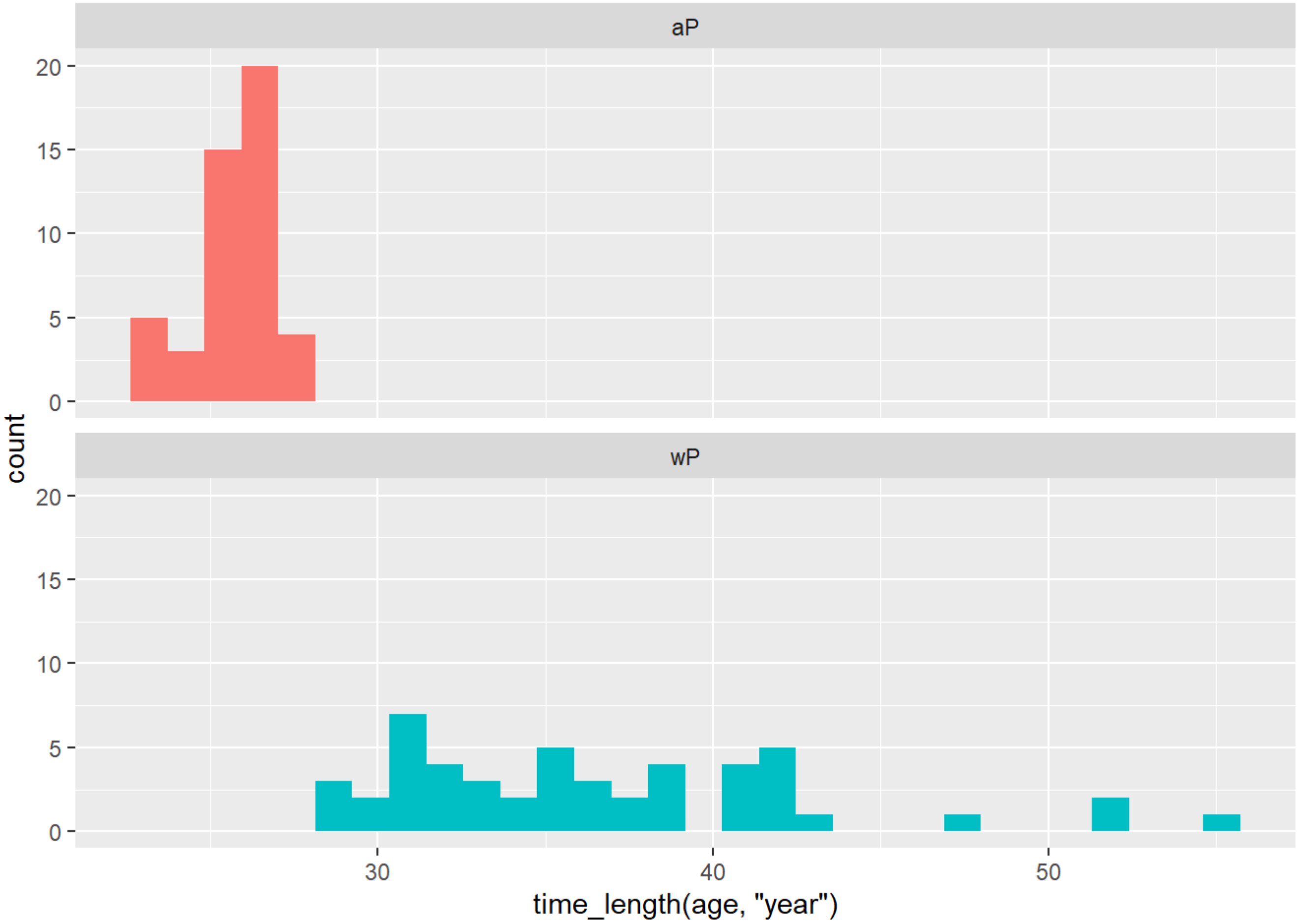
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)

age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

> 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`.
```



Now let's read some database tables from CMI-PB:

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", 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
```

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", 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
```

I want to "join" (a.k.a. "merge"/link/etc.) the `subject` and `specimen` tables together. I wil use the **dplyr** package.

```
library(dplyr)

meta <- inner_join(x = subject, y = specimen)
head(meta)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
```

```
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 13670 days           1
2    1986-01-01    2016-09-12 2020_dataset 13670 days           2
3    1986-01-01    2016-09-12 2020_dataset 13670 days           3
4    1986-01-01    2016-09-12 2020_dataset 13670 days           4
5    1986-01-01    2016-09-12 2020_dataset 13670 days           5
6    1986-01-01    2016-09-12 2020_dataset 13670 days           6
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                          736                           736         Blood
3                            1                             1         Blood
4                            3                             3         Blood
5                            7                             7         Blood
6                           11                            14         Blood
  visit
1     1
2    10
3     2
4     3
5     4
6     5
```

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

```
  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
```

Now I can join this data with the “meta” data we created.

> 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.

```r
abdata <- inner_join(x = meta, y = ab)
dim(abdata)
```

```
[1] 32675    21
```

> 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 
```

There are way less specimens for visit 8 because this project is still ongoing and we have not got that data for all individuals yet.

Examine IgG1 Ab Titer levels

We will use the `filter()` function from dplyr to focus on just IgG1 isotype and visits 1 to 7. Exclude visit 8 because there isn't much data reported.

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

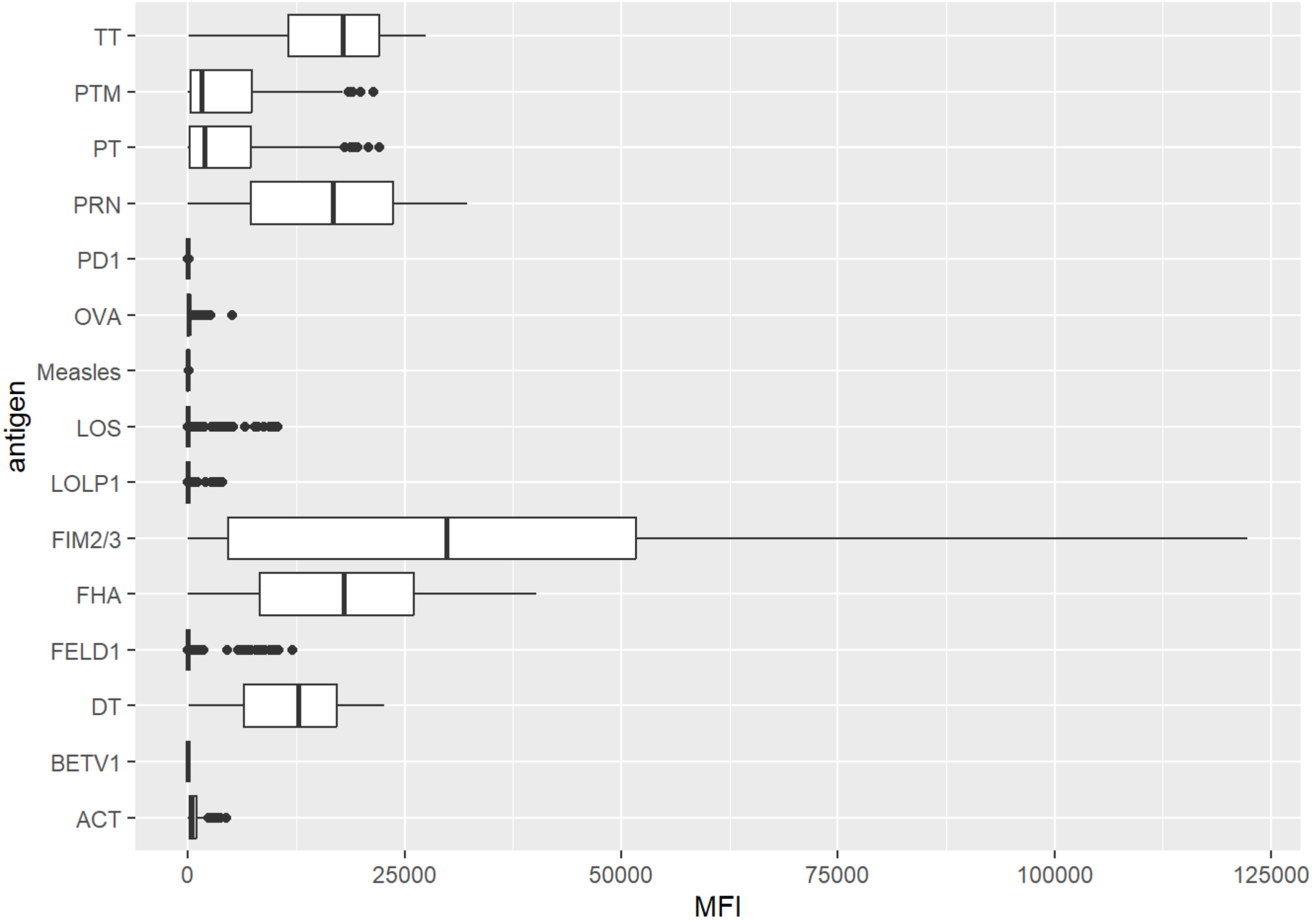
```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 13670 days           1
2    1986-01-01    2016-09-12 2020_dataset 13670 days           1
3    1986-01-01    2016-09-12 2020_dataset 13670 days           1
4    1986-01-01    2016-09-12 2020_dataset 13670 days           1
```

```
5    1986-01-01    2016-09-12 2020_dataset 13670 days           1
6    1986-01-01    2016-09-12 2020_dataset 13670 days           1
  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 isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1    IgG1                TRUE     ACT 274.355068      0.6928058 IU/ML
2     1    IgG1                TRUE     LOS  10.974026      2.1645083 IU/ML
3     1    IgG1                TRUE   FELD1   1.448796      0.8080941 IU/ML
4     1    IgG1                TRUE   BETV1   0.100000      1.0000000 IU/ML
5     1    IgG1                TRUE   LOLP1   0.100000      1.0000000 IU/ML
6     1    IgG1                TRUE Measles  36.277417      1.6638332 IU/ML
  lower_limit_of_detection
1                 3.848750
2                 4.357917
3                 2.699944
4                 1.734784
5                 2.550606
6                 4.438966
```

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

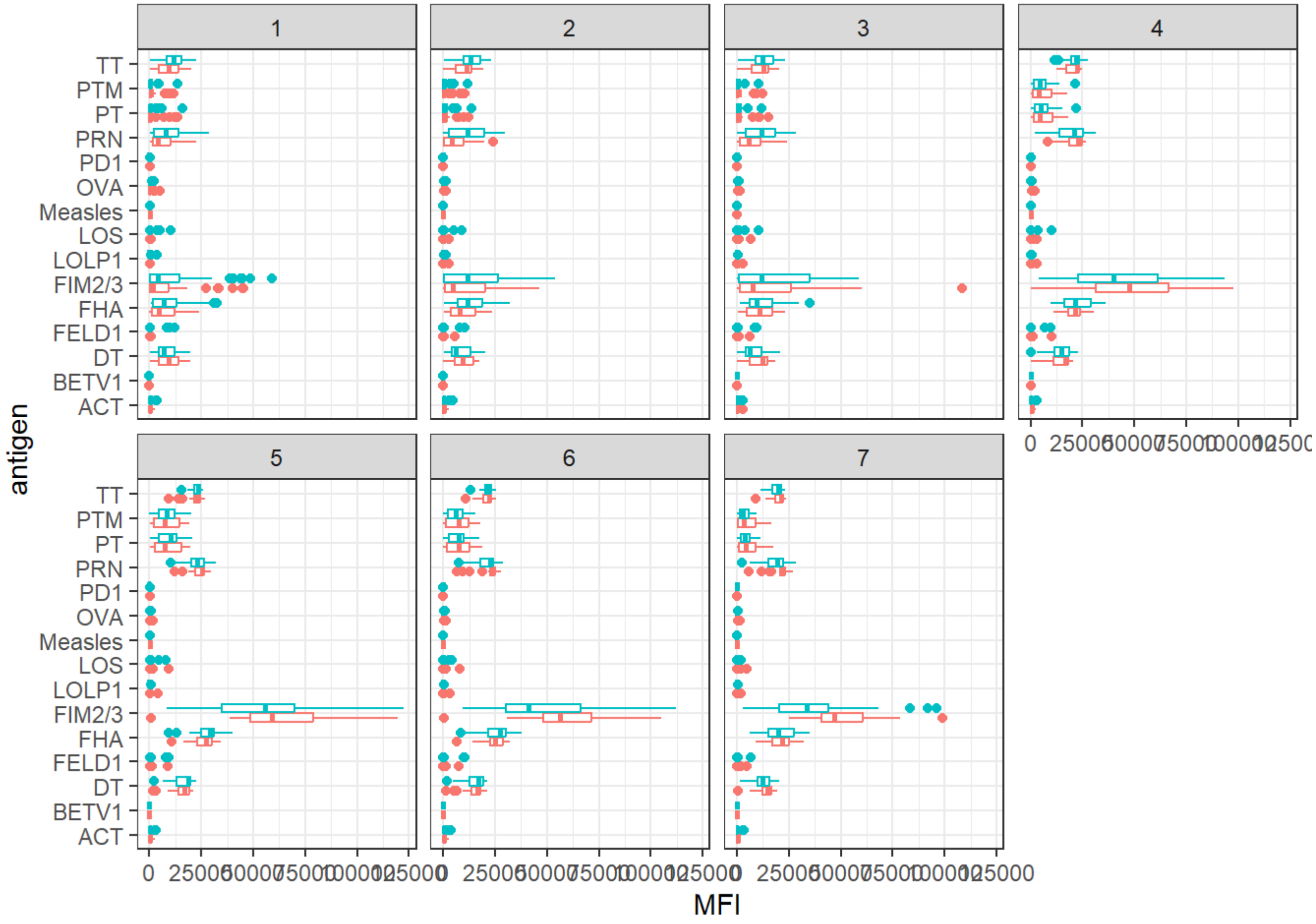
Here’s a boxplot of antigen levels over time.

```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot()
```



And let’s facet this by visit:

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



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

Clearly FIM2/3 changes over time. This is “fimbrial protein” that makes the bacterial pilus involved in cell adhesion. This likely is an identifiable tag by antibodies and causes a immune response compared to others.

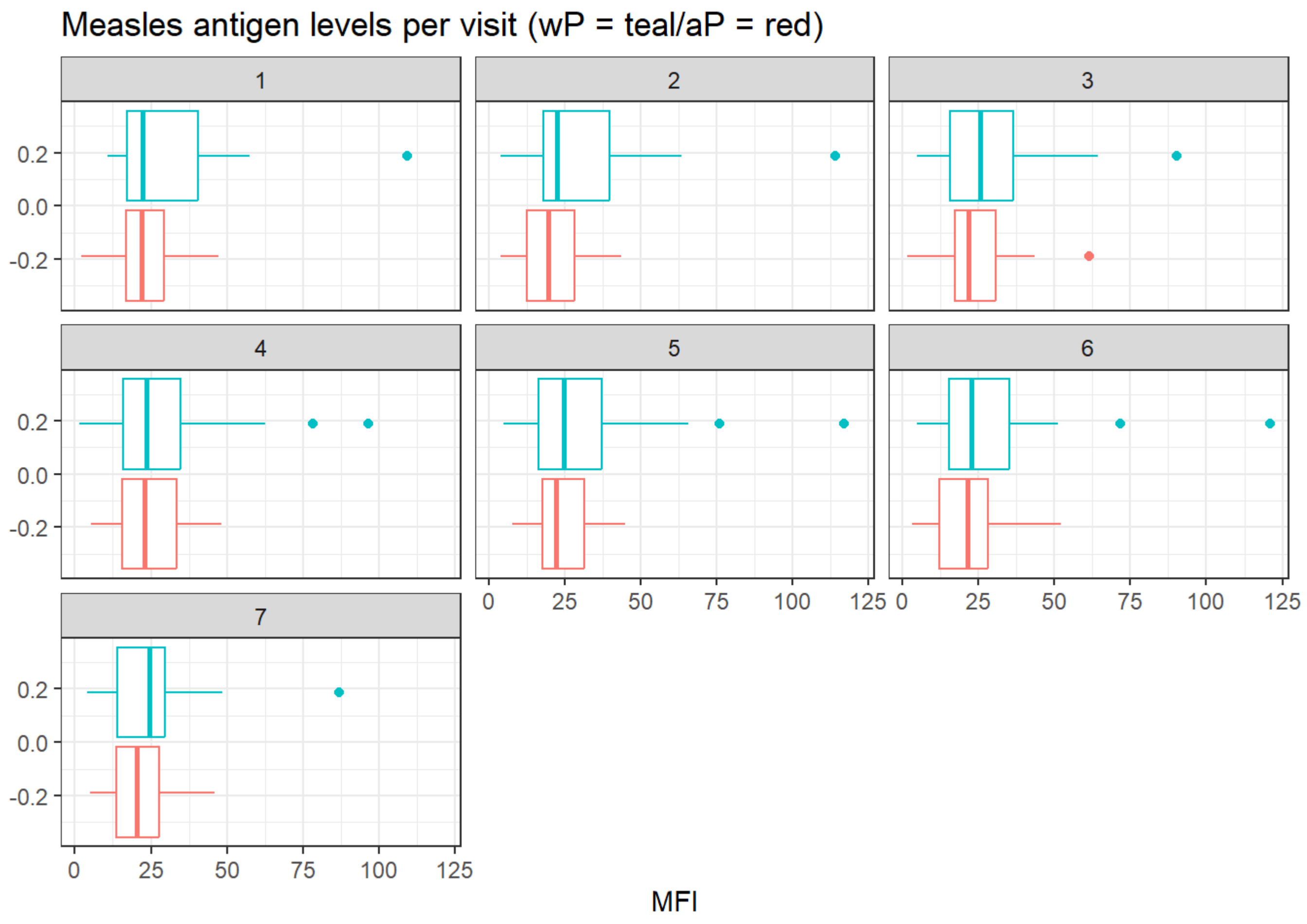
PT - Pertussis

FHAB - Filamentous hemagglutinin

> 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).

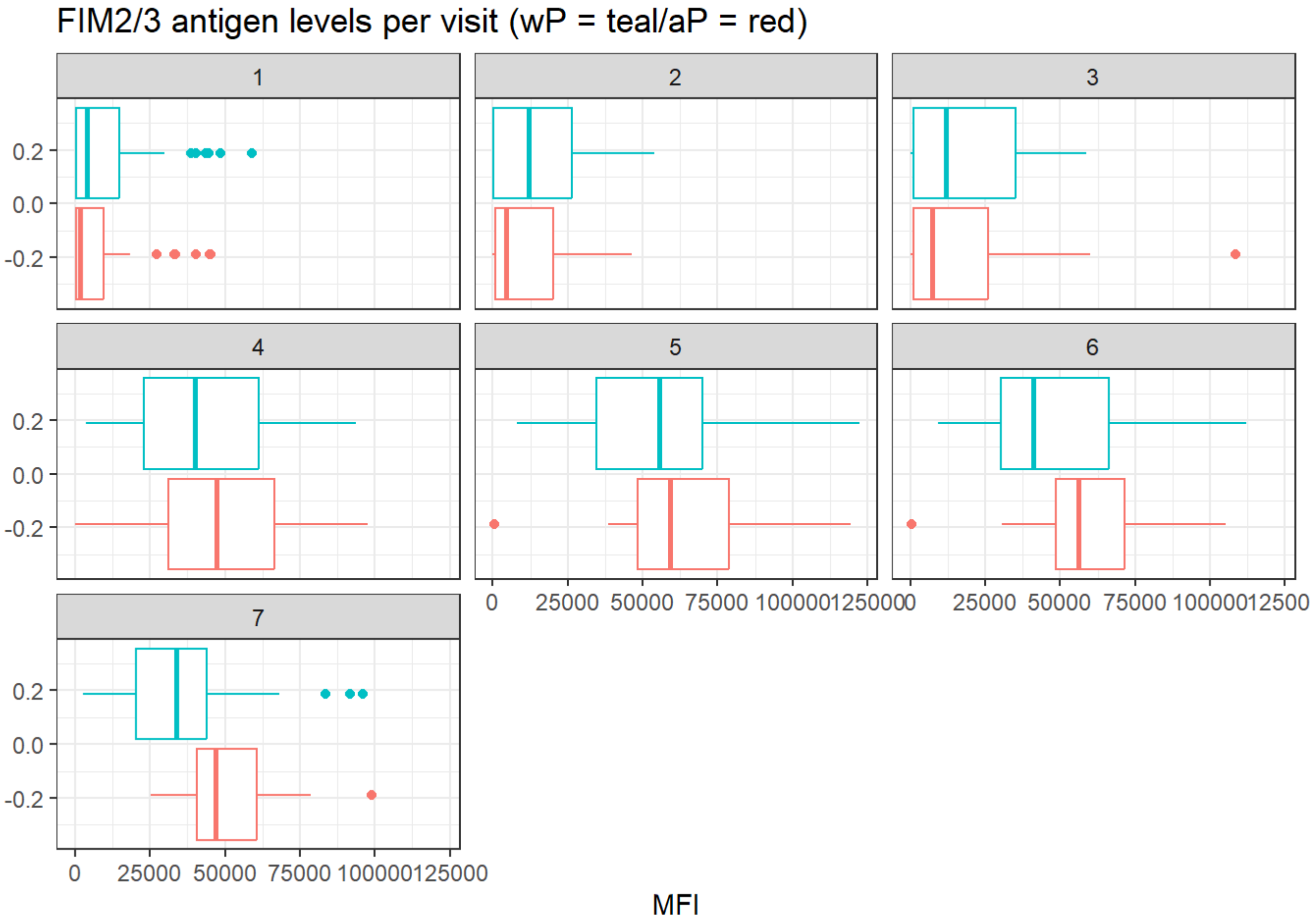
Filter for Measles:

```
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 for FIM2/3:

```
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 levels clearly rise over time and far exceed those of Measles. They also appear to peak at visit 5 and then decline. This trend appears similar for for wP and aP subjects.

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

It's hard to tell the clear differences between these responses as it canbe subject to anlaysis and these datasets can also change unknowingly over time.

Obtaining CMI-PB RNASeq data

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 use the following URL

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

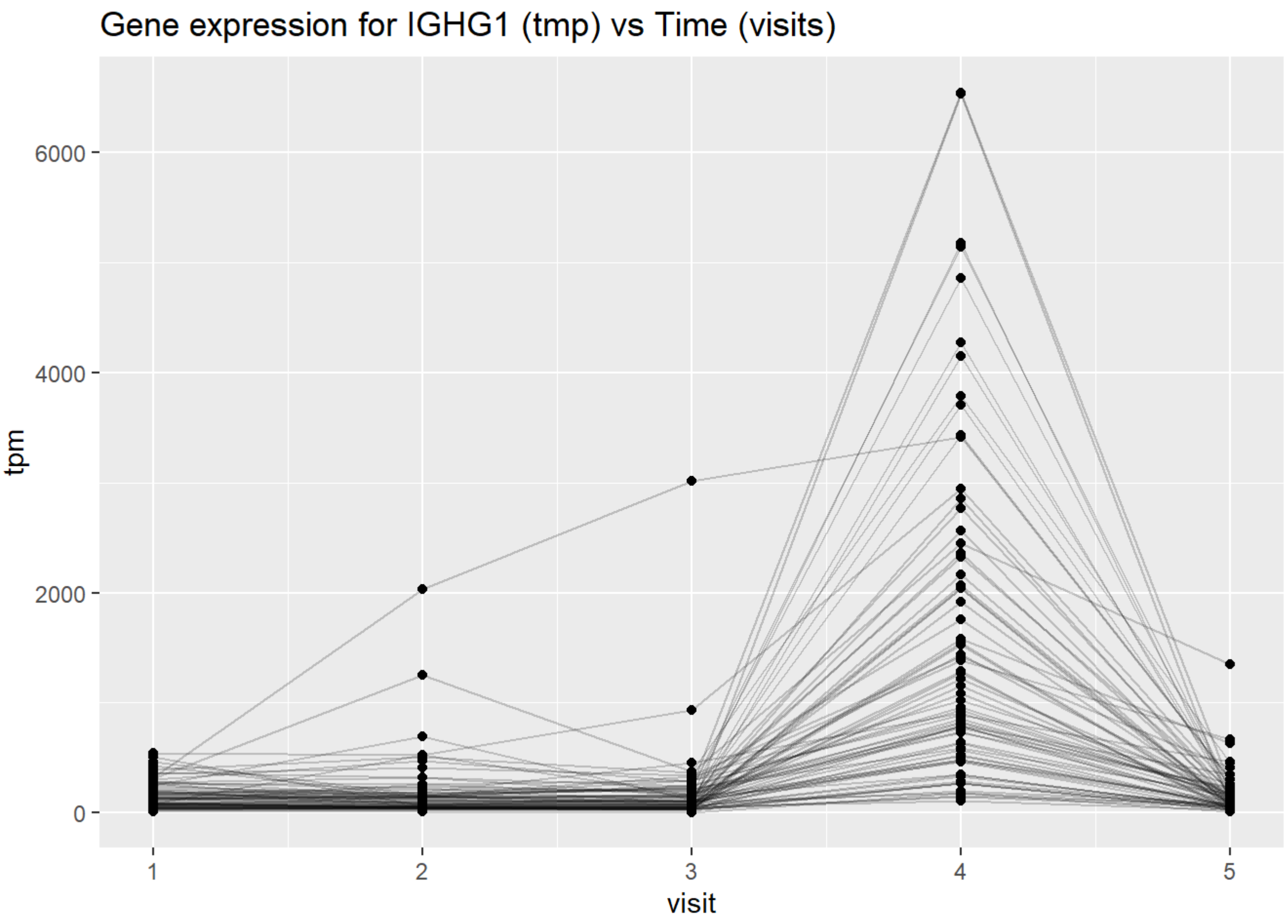
The link above is for the key gene involved in expressing any IgG1 antibody, namely the IGHG1 gene. Let's read available RNA-Seq data for this gene into R and investigate the time course of it's gene expression values.

To facilitate further analysis we need to "join" the rna expression data with our metadata meta.

```
#we already joined our meta data previously: meta <- inner_join(specimen, subject)
ssrna <- inner_join(rna, meta)
```

> 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) +
  ggtitle("Gene expression for IGHG1 (tmp) vs Time (visits)")
```



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

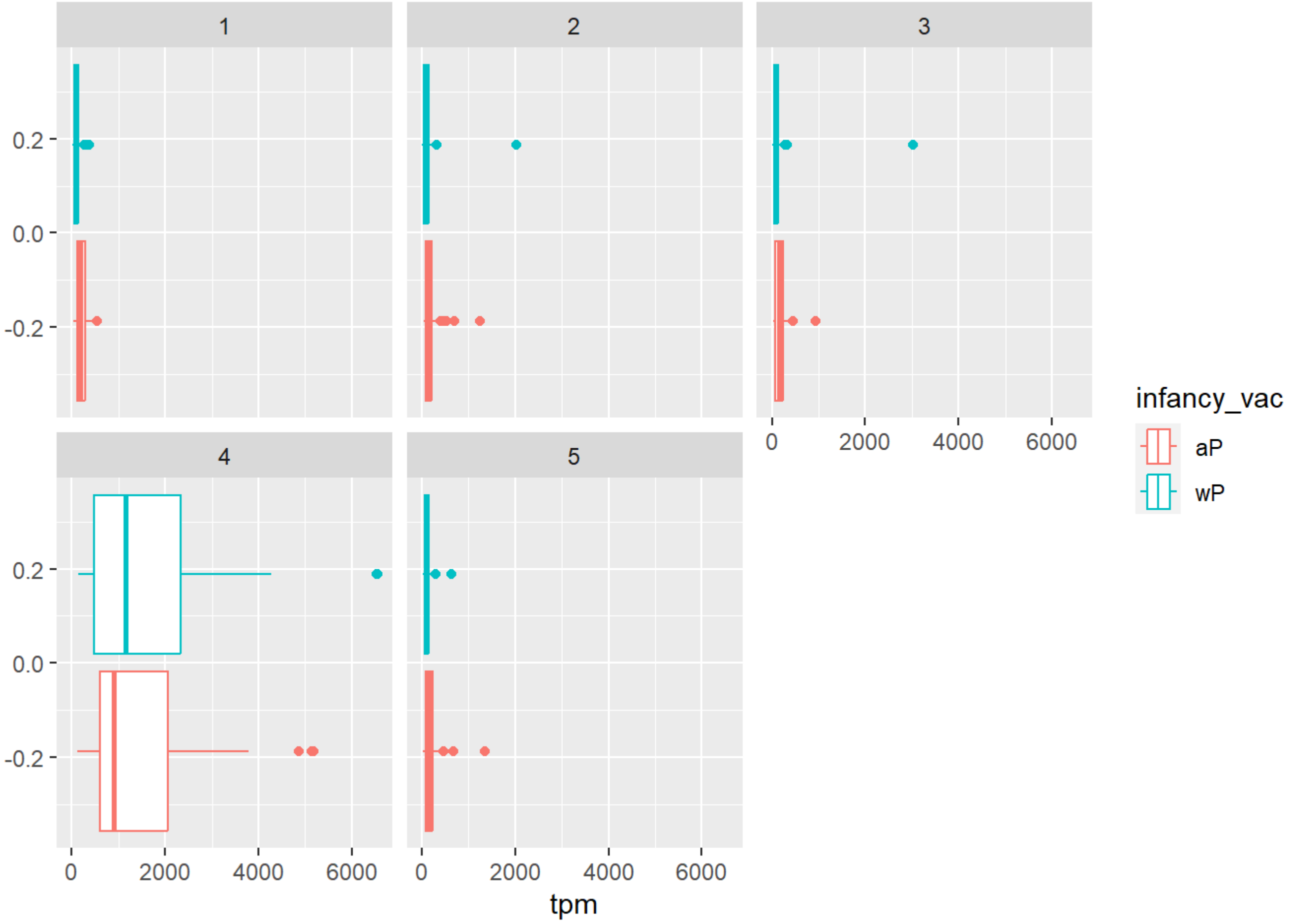
The expression of this IGHG1 gene is at its maximum level during the 4th visit.

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

Yes this pattern matches the trend of antibody data because cells make long-lived antibodies.

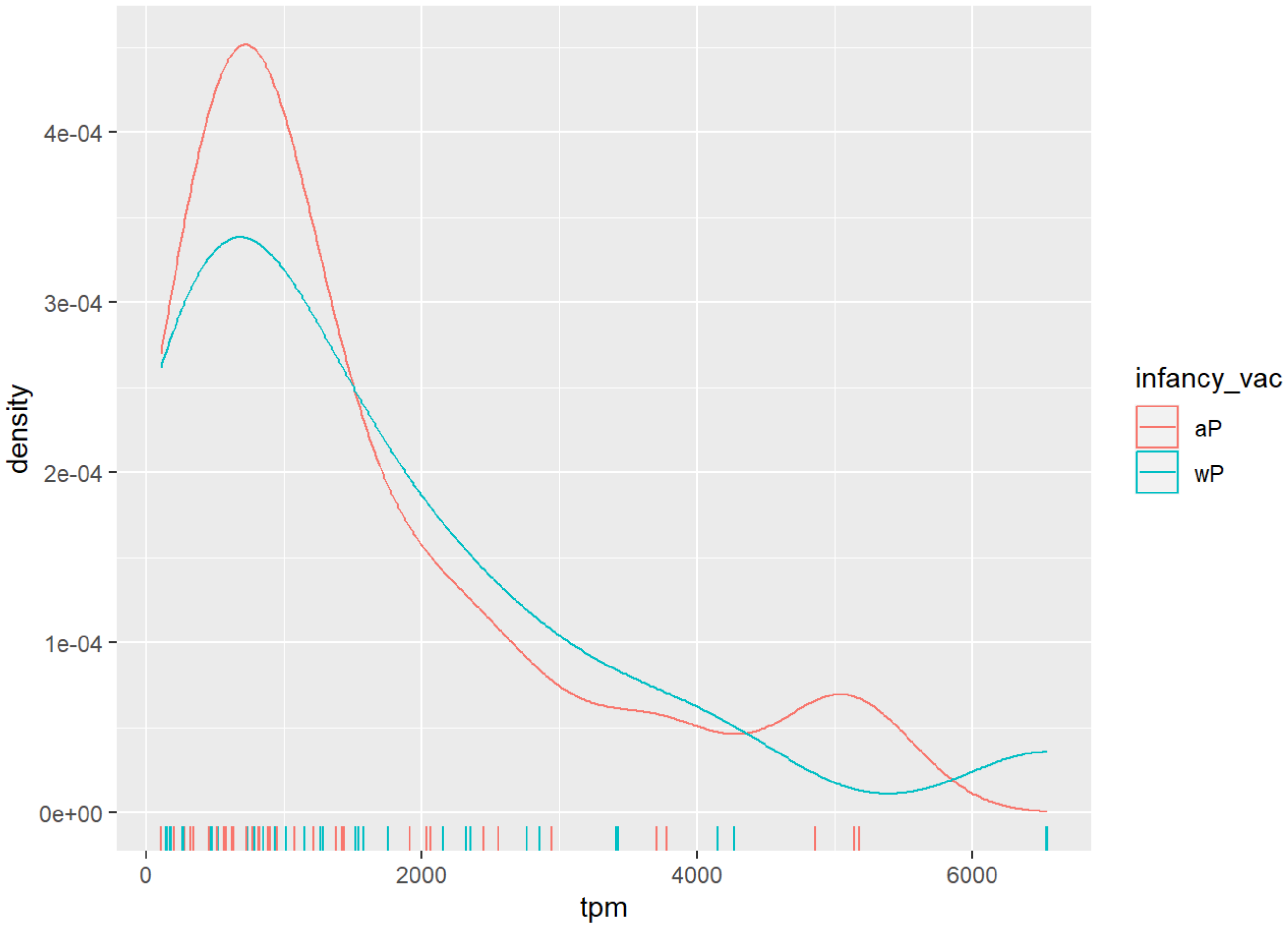
We can dig deeper and color and/or facet by infancy_vac status:

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



Even if we focus on a particular visit, there is no obvious wP vs. aP differences.

```
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