pertussis mini project

Liz Chamiec-Case

Investigating pertussis cases by year

```r
library(datapasta)
```

```
Warning: package 'datapasta' was built under R version 4.2.2
```

```r
cdc <- tibble::tribble(
                               ~Year,    ~Reported_Cases,
                                1922,   107473,
                                    1923,   164191,
                                    1924,   165418,
                                    1925,   152003,
                                    1926,   202210,
                                    1927,   181411,
                                    1928,   161799,
                                    1929,   197371,
                                    1930,   166914,
                                    1931,   172559,
                                    1932,   215343,
                                    1933,   179135,
                                    1934,   265269,
                                    1935,   180518,
                                    1936,   147237,
                                    1937,   214652,
                                    1938,   227319,
                                    1939,   103188,
                                    1940,   183866,
                                    1941,   222202,
                                    1942,   191383,
                                    1943,   191890,
1944,   109873,
1945,   133792,
1946,   109860,
1947,   156517,
1948,   74715,
1949,   69479,
1950,   120718,
1951,   68687,
1952,   45030,
1953,   37129,
1954,   60886,
1955,   62786,
1956,   31732,
1957,   28295,
1958,   32148,
1959,   40005,
1960,   14809,
1961,   11468,
1962,   17749,
1963,   17135,
1964,   13005,
1965,   6799,
1966,   7717,
1967,   9718,
1968,   4810,
1969,   3285,
1970,   4249,
1971,   3036,
1972,   3287,
1973,   1759,
1974,   2402,
1975,   1738,
1976,   1010,
1977,   2177,
1978,   2063,
1979,   1623,
1980,   1730,
1981,   1248,
1982,   1895,
1983,   2463,
1984,   2276,
                                    1985,   3589,
                                    1986,   4195,
                                    1987,   2823,
                                    1988,   3450,
                                    1989,   4157,
                                    1990,   4570,
                                    1991,   2719,
                                    1992,   4083,
                                    1993,   6586,
                                    1994,   4617,
                                    1995,   5137,
                                    1996,   7796,
                                    1997,   6564,
                                    1998,   7405,
                                    1999,   7298,
                                    2000,   7867,
                                    2001,   7580,
                                    2002,   9771,
                                    2003,   11647,
                                    2004,   25827,
                                    2005,   25616,
                                    2006,   15632,
                                    2007,   10454,
                                    2008,   13278,
                                    2009,   16858,
                                    2010,   27550,
                                    2011,   18719,
                                    2012,   48277,
                                    2013,   28639,
                                    2014,   32971,
                                    2015,   20762,
                                  2016, 17972,
                                  2017, 18975,
                                  2018, 15609,
                                  2019, 18617
    )

cdc
```

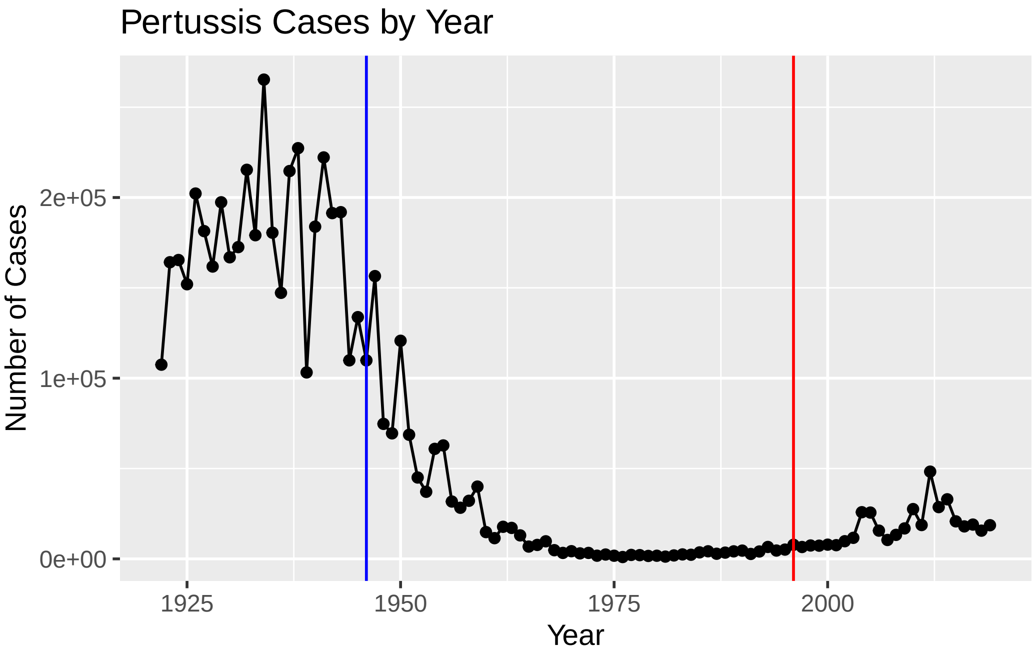
```
# A tibble: 98 x 2
    Year Reported_Cases
   <dbl>          <dbl>
```

```
 1  1922         107473
 2  1923         164191
 3  1924         165418
 4  1925         152003
 5  1926         202210
 6  1927         181411
 7  1928         161799
 8  1929         197371
 9  1930         166914
10  1931         172559
# ... with 88 more rows
```

```r
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.2
```

```r
ggplot(cdc) +
  aes(x=Year, y=Reported_Cases) +
  geom_point() +
  geom_line() +
  labs(x="Year",y="Number of Cases",title="Pertussis Cases by Year") +
  geom_vline(xintercept=1946,color="blue") +
  geom_vline(xintercept=1996,color="red")
```



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 was introduced, the number of cases started to rise for the first time in years. This could be due to distrust in a new vaccine, development of new strains that the newer vaccines are less effective against, or better testing methods.

```
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.2.2
```

```
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 many 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
```

```
library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.2.2
```

```
Loading required package: timechange
```

```
Warning: package 'timechange' was built under R version 4.2.2


Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
# i) average age of wP individuals
wP <- subset(subject,infancy_vac == "wP")
wP['age_days'] <- today() - ymd(wP$year_of_birth)
wP['age_years'] <- round(time_length( today() - ymd(wP$year_of_birth),  "years"))
summary(wP$age_years)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.00   32.00   35.00   36.16   40.00   55.00
```

```
# ii) average age of aP individuals
aP <- subset(subject,infancy_vac == "aP")
aP['age_days'] <- today() - ymd(aP$year_of_birth)
aP['age_years'] <- round(time_length( today() - ymd(aP$year_of_birth),  "years"))
summary(aP$age_years)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   25.00   26.00   25.32   26.00   27.00
```

iii) are they significantly different?

In the absence of a statistical test, we can tell that the average ages are likely significantly different because there is no overlap in the data (max(aP) < min(wP)). This makes sense since the wP vaccine was likely discontinued when the aP vaccine was introduced.

also significant with a wilcox test:

```
x <- t.test(wP$age_years,aP$age_years)

x$p.value
```

```
[1] 1.315544e-16
```

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

```
subject['age_at_boost'] <- round(time_length(ymd(subject$date_of_boost) - ymd(subject$year
subject['age_at_boost']
```

```
   age_at_boost
1            31
2            51
3            34
4            29
5            26
6            29
7            36
8            34
9            21
10           35
11           31
12           35
13           20
14           24
15           28
16           30
17           37
18           20
19           23
20           32
21           26
22           24
23           26
24           29
25           43
26           47
27           47
28           29
29           21
30           21
31           28
32           24
33           24
34           21
```

35	21
36	31
37	26
38	32
39	27
40	26
41	21
42	20
43	22
44	19
45	21
46	19
47	19
48	22
49	20
50	21
51	19
52	23
53	20
54	21
55	19
56	36
57	34
58	32
59	26
60	25
61	29
62	34
63	20
64	35
65	20
66	29
67	28
68	20
69	27
70	34
71	26
72	20
73	19
74	20
75	32
76	23
77	32

```
78              20
79              19
80              19
81              20
82              19
83              21
84              19
85              20
86              20
87              20
88              19
89              19
90              20
91              20
92              20
93              21
94              20
95              20
96              20
```

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

Yes - aP has a much smaller mean and variance whereas wP has a higher mean and variance. There is almost no overlap between the two groups.

```
# Or use wilcox.test()
x <- t.test(wP$age_years,aP$age_years)

x$p.value
```

```
[1] 1.315544e-16
```

Joining multiple tables

```
# 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)

# join datasets
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
```

```r
meta <- full_join(specimen, subject)
```

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

```r
dim(meta)
```

```
[1] 729  14
```

```r
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_at_boost
1           31
2           31
3           31
4           31
5           31
6           31
```

```
# join meta and titer
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
[1] 32675    21
```

```
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?

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

```

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

There are substantially fewer visit 8 specimens as compared to the other visits

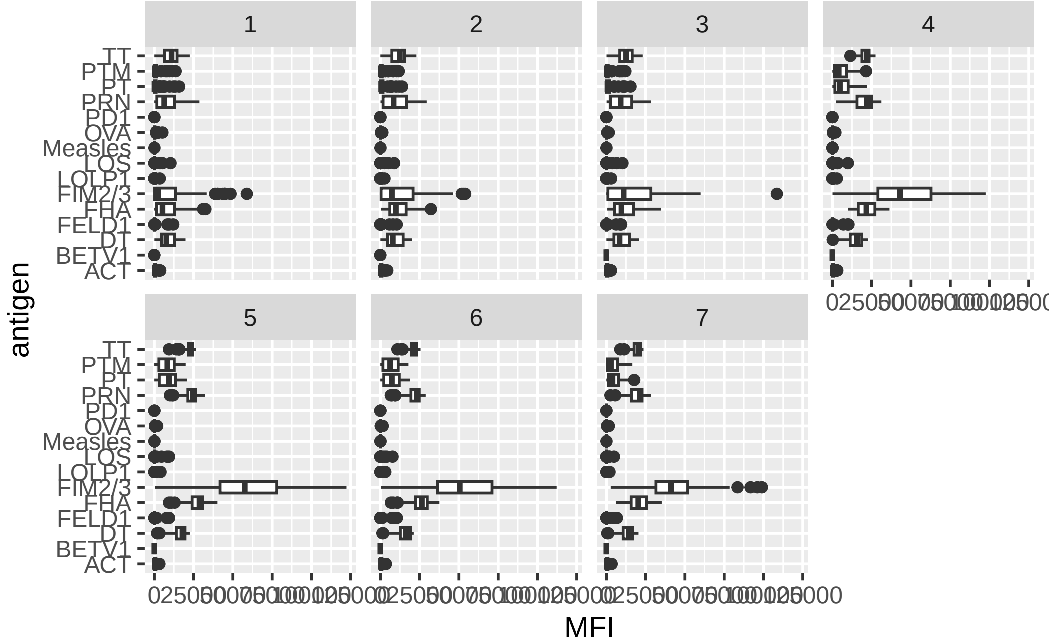
Examine IgG1 Ab titer levels

```
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_at_boost
1           31
2           31
3           31
4           31
5           31
6           31
```

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

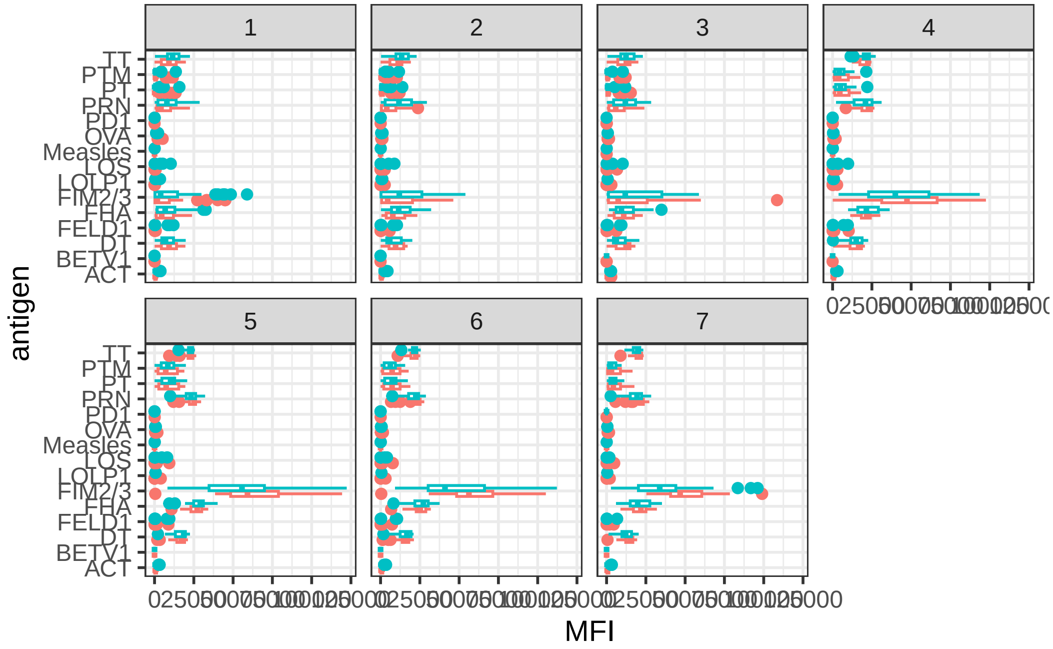
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



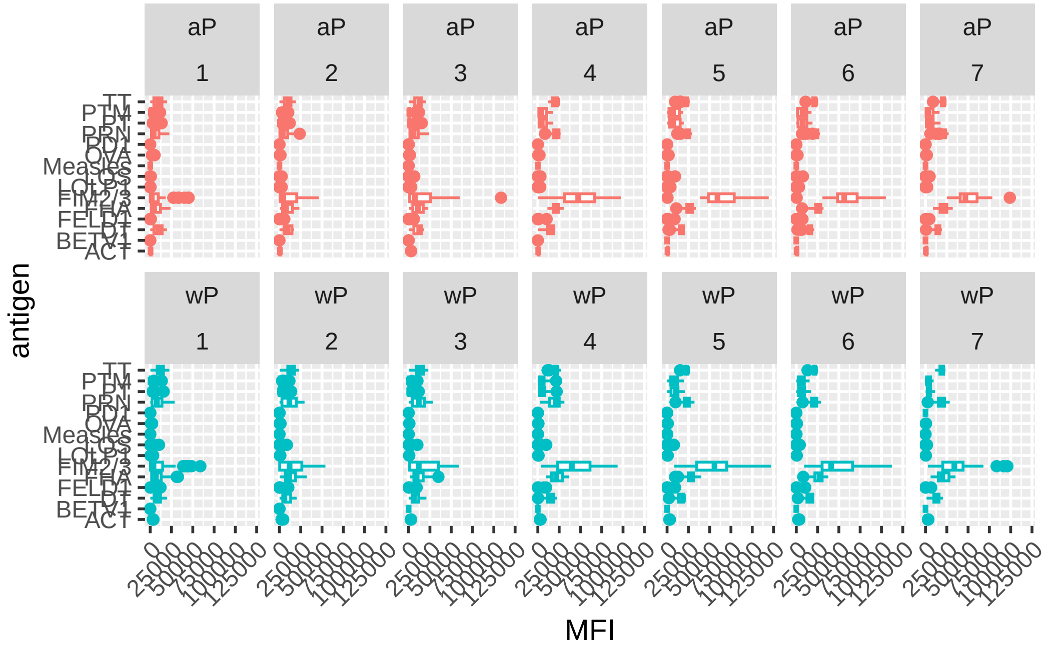
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3 differs quite a bit over time, although it seems to have a lot of outliers. It also saturates the plot, so it's difficult to tell if other antigens are responding similarly. This trend is not well reflected in the CBI-PM data, though.

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

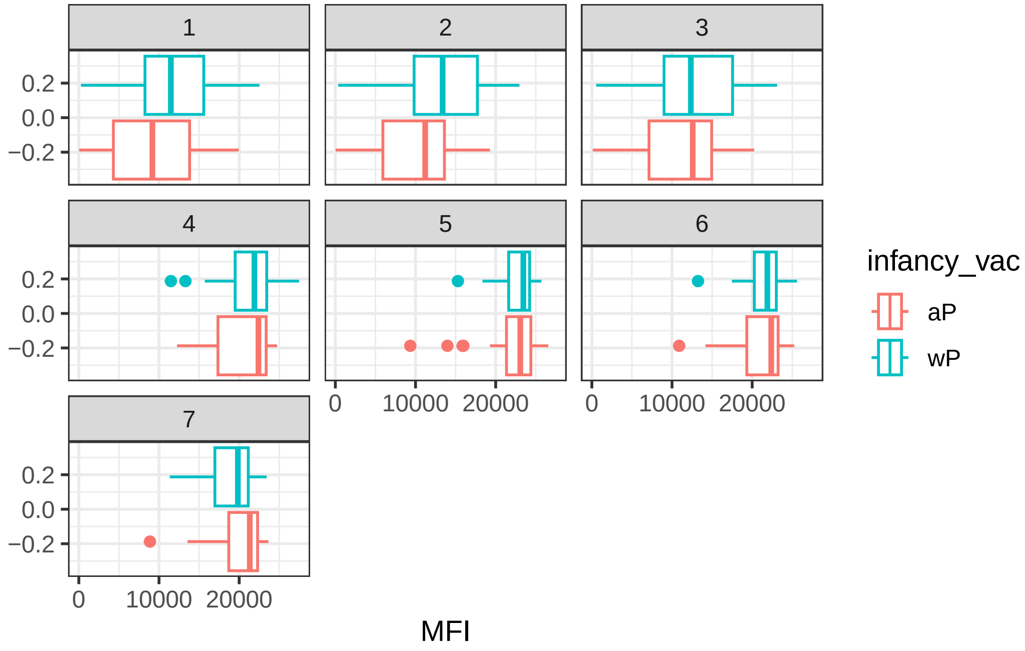


```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```

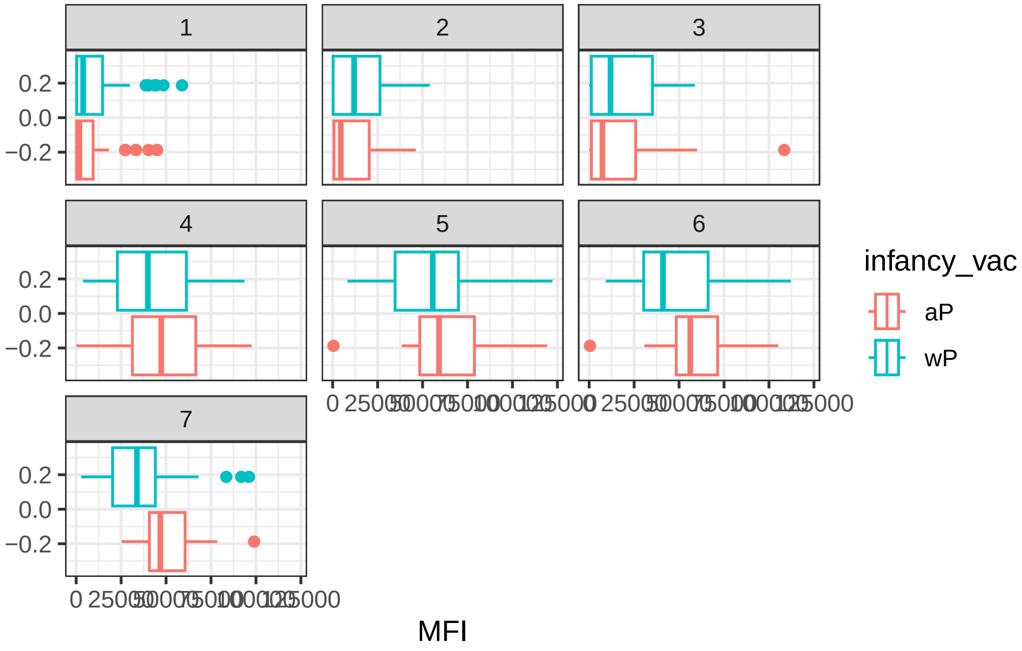


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate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

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



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



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

Both TT and FIM2/3 antigen time courses rise over time, perhaps TT more so than FIM2/3 but difficult to tell because TT has large error bars. FIM2/3 seems to peak around visit 5 whereas TT does not come back down.

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

In FIM2/3, aP seems to increase more than wP does over time. The effect is smaller if present at all in TT.

Obtaining CMI-PB RNASeq data

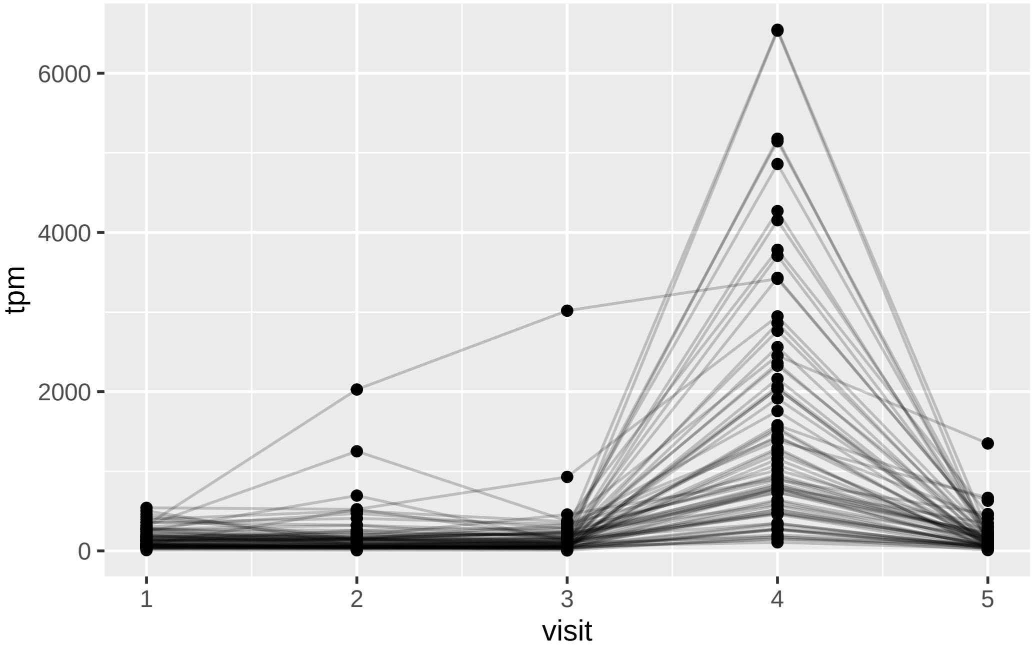
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.

rna <- read_json(url, simplifyVector = TRUE)

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

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

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



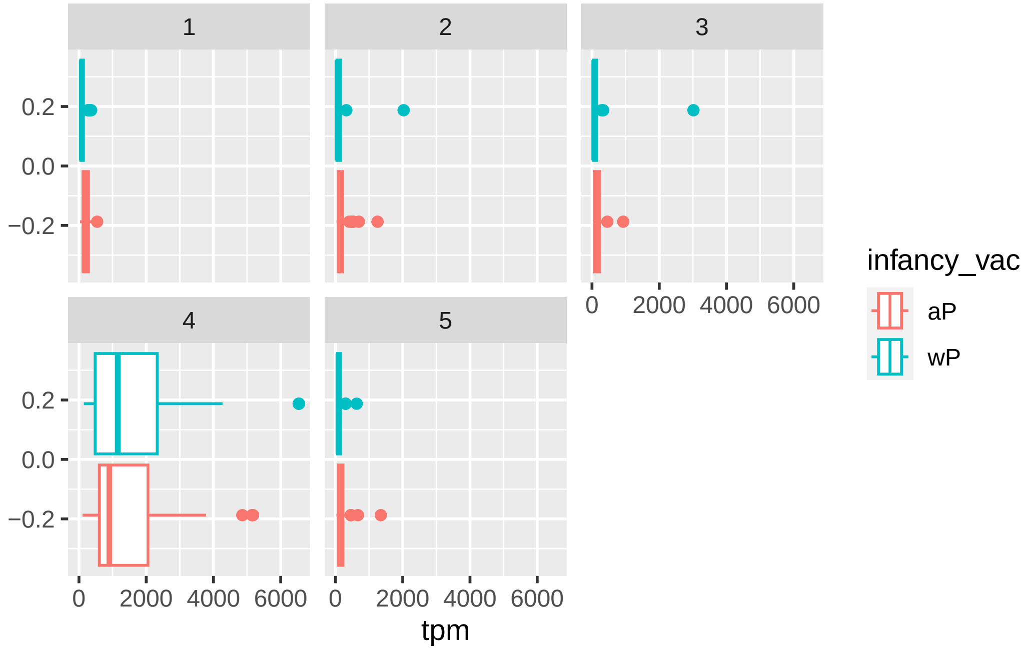
Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The expression of this gene peaks very specifically at visit 4 by more than double the expression at any other visits.

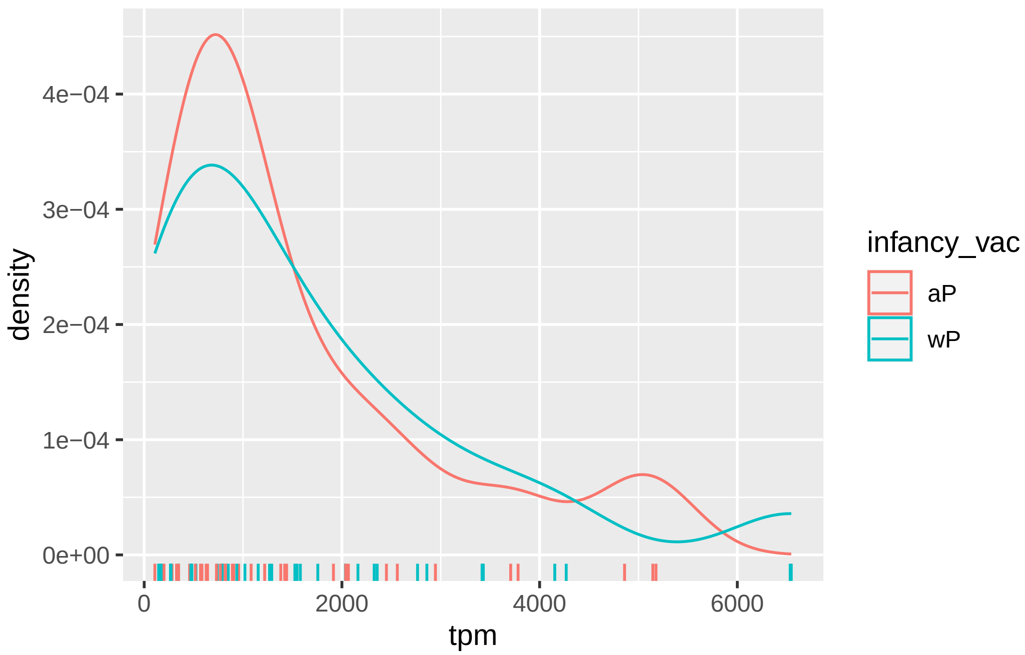
Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

This makes sense - the antibody titer data for FIM2/3 peaks at visit 5 then declines after that. It takes time from expression of the gene to produce/circulate the protein so, depending on how far apart the visits are, we might expect to see this.

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



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


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