class15

Phoebe LI

3/9/2022

Dr. Bjoern Peters project

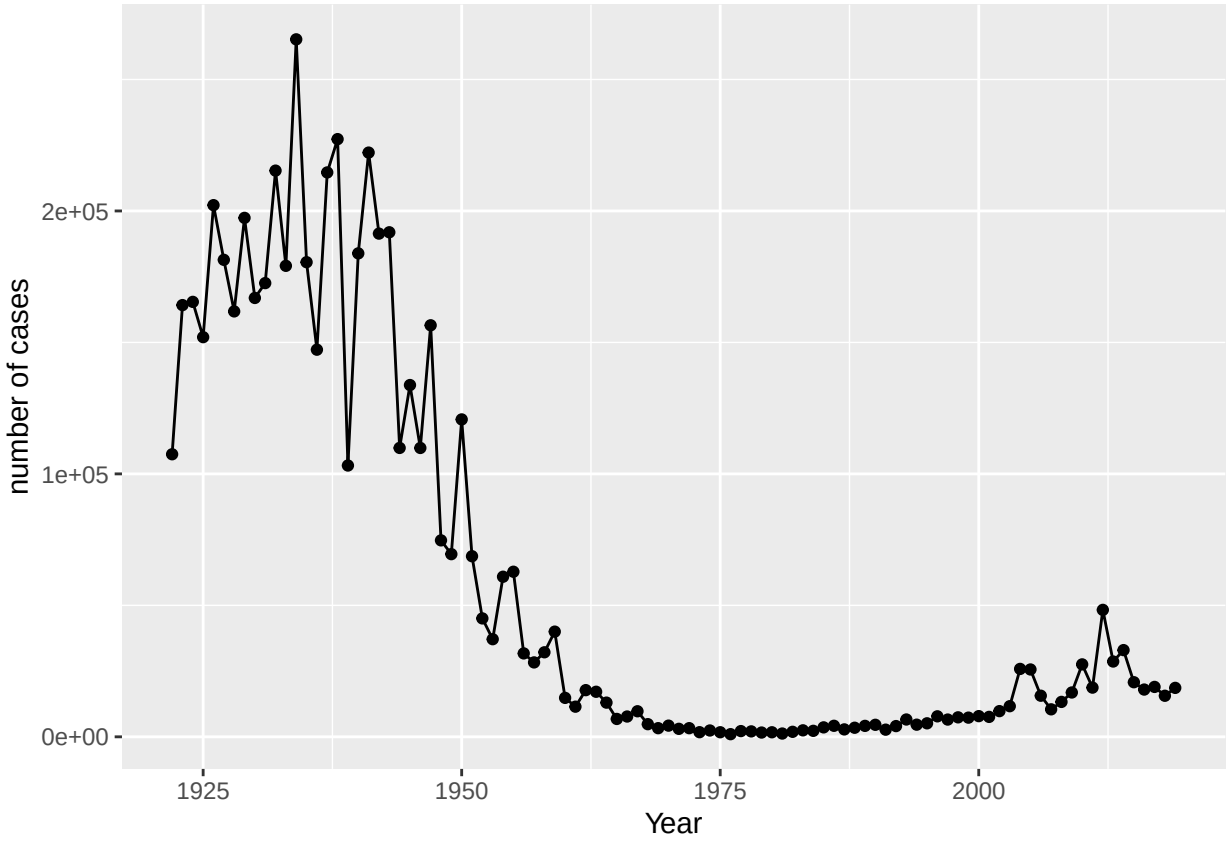
1. Investigating Pertussis Resurgence

Here we used Addins from datapasta I get data from https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

```
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,
```

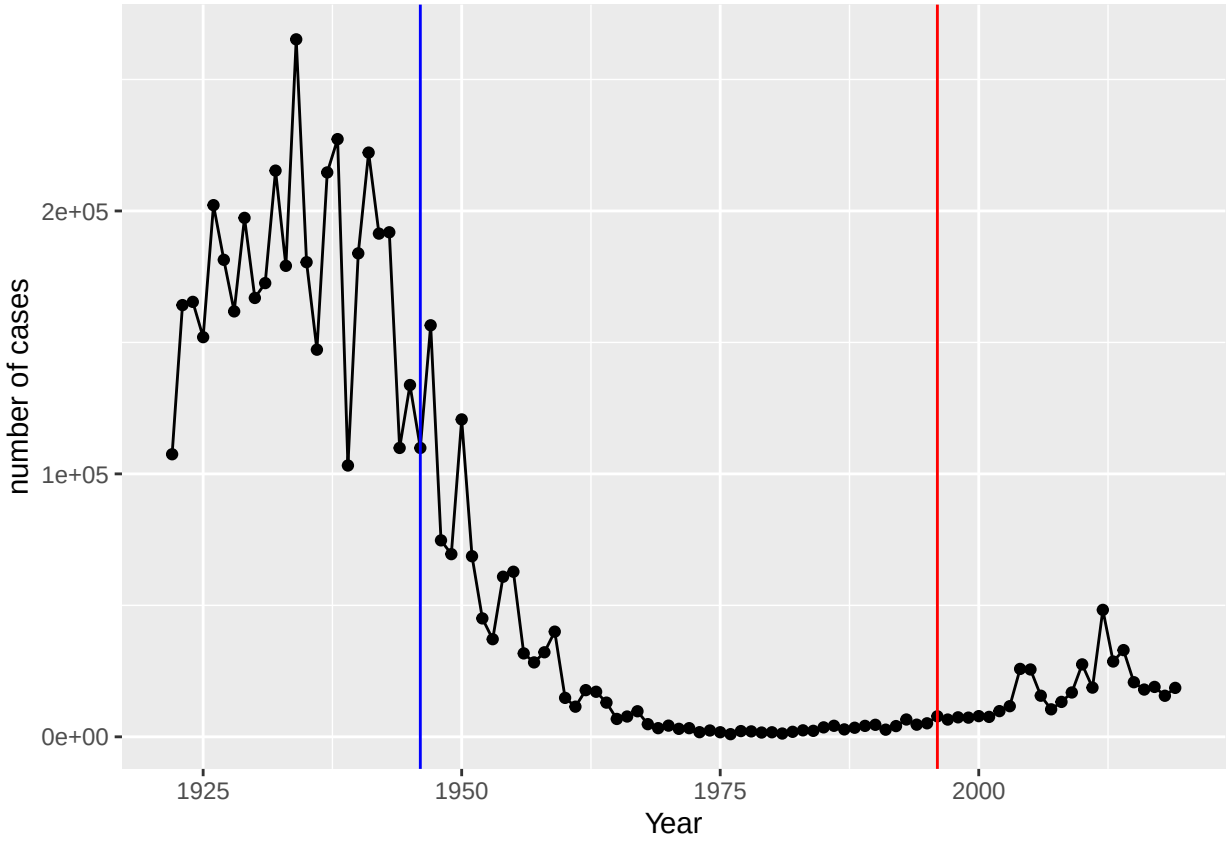
```
                                                    214652,227319,103188,183866,
                                                    222202,191383,191890,
                                                    109873,133792,109860,
                                                    156517,74715,69479,120718,
                                                    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(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x="Year", y= "number of cases")
```



2. A tale of two vaccines (wP & aP)

```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x="Year", y= "number of cases")+
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?

Can due to low vaccination rate and virus mutations. Also there are more traveling population.

They are 10 years old. They are the first generation using the new vaccine.

3. Exploring CMI-PB data

```
# Allows us to read, write and process JSON data
library(jsonlite)
```

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

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

```
head(subject, 3)
```

```
##   subject_id infancy_vac biological_sex              ethnicity  race
## 1          1          wP         Female Not Hispanic or Latino White
## 2          2          wP         Female Not Hispanic or Latino White
## 3          3          wP         Female                 Unknown White
##   year_of_birth date_of_boost   study_name
```

```
## 1    1986-01-01    2016-09-12 2020_dataset
## 2    1968-01-01    2019-01-28 2020_dataset
## 3    1983-01-01    2016-10-10 2020_dataset
```

Q4. How 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
```

Side-Note: Working with dates

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
today()
```

```
## [1] "2022-03-09"
```

```
today() - ymd("2000-01-01")
```

```
## Time difference of 8103 days
```

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

```
## [1] 22.1848
```

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_now<- time_length( today() - ymd(subject$year_of_birth),  "years")
age_now
```

```
##  [1] 36.18344 54.18480 39.18412 34.18480 31.18412 34.18480 41.18275 37.18275
##  [9] 26.18480 40.18344 36.18344 40.18344 25.18275 29.18275 33.18275 35.18412
## [17] 42.18480 25.18275 28.18344 35.18412 29.18275 27.18412 29.18275 32.18344
## [25] 46.18480 50.18480 50.18480 32.18344 24.18344 24.18344 31.18412 27.18412
## [33] 27.18412 24.18344 24.18344 34.18480 29.18275 35.18412 30.18480 29.18275
## [41] 24.18344 23.18412 25.18275 22.18480 24.18344 22.18480 22.18480 25.18275
## [49] 23.18412 24.18344 22.18480 26.18480 23.18412 24.18344 22.18480 41.18275
## [57] 39.18412 37.18275 31.18412 30.18480 34.18480 39.18412 25.18275 40.18344
## [65] 25.18275 34.18480 33.18275 25.18275 32.18344 39.18412 31.18412 25.18275
## [73] 24.18344 25.18275 37.18275 28.18344 37.18275 25.18275 24.18344 24.18344
## [81] 25.18275 24.18344 26.18480 24.18344 25.18275 25.18275 25.18275 24.18344
## [89] 24.18344 25.18275 25.18275 25.18275 26.18480 25.18275 25.18275 25.18275
```

```
avegae_age<-mean(age_now)
avegae_age
```

```
## [1] 30.03779
```

```
# for ap and wp average age
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 <- subject %>% filter(infancy_vac == "aP")
wp <- subject %>% filter(infancy_vac == "wP")
```

```
ap_age_now<- time_length( today() - ymd(ap$year_of_birth),  "years")
wp_age_now<- time_length( today() - ymd(wp$year_of_birth),  "years")
ap_age_now
```

```
##  [1] 26.18480 25.18275 25.18275 24.18344 24.18344 24.18344 24.18344 24.18344
##  [9] 23.18412 25.18275 22.18480 24.18344 22.18480 22.18480 25.18275 23.18412
## [17] 24.18344 22.18480 26.18480 23.18412 24.18344 22.18480 25.18275 25.18275
## [25] 25.18275 25.18275 24.18344 25.18275 25.18275 24.18344 24.18344 25.18275
## [33] 24.18344 26.18480 24.18344 25.18275 25.18275 25.18275 24.18344 24.18344
## [41] 25.18275 25.18275 25.18275 26.18480 25.18275 25.18275 25.18275
```

```
wp_age_now
```

```
##  [1] 36.18344 54.18480 39.18412 34.18480 31.18412 34.18480 41.18275 37.18275
##  [9] 40.18344 36.18344 40.18344 29.18275 33.18275 35.18412 42.18480 28.18344
## [17] 35.18412 29.18275 27.18412 29.18275 32.18344 46.18480 50.18480 50.18480
## [25] 32.18344 31.18412 27.18412 27.18412 34.18480 29.18275 35.18412 30.18480
## [33] 29.18275 41.18275 39.18412 37.18275 31.18412 30.18480 34.18480 39.18412
## [41] 40.18344 34.18480 33.18275 32.18344 39.18412 31.18412 37.18275 28.18344
## [49] 37.18275
```

```
ap_avegae_age<-mean(ap_age_now)
wp_avegae_age<-mean(wp_age_now)
ap_avegae_age
```

```
## [1] 24.5026
```

```
wp_avegae_age
```

```
## [1] 35.34705
```

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
```

```
age_boost<- time_length( ymd(subject$date_of_boost) - ymd(subject$year_of_birth),  "years")
age_boost
```

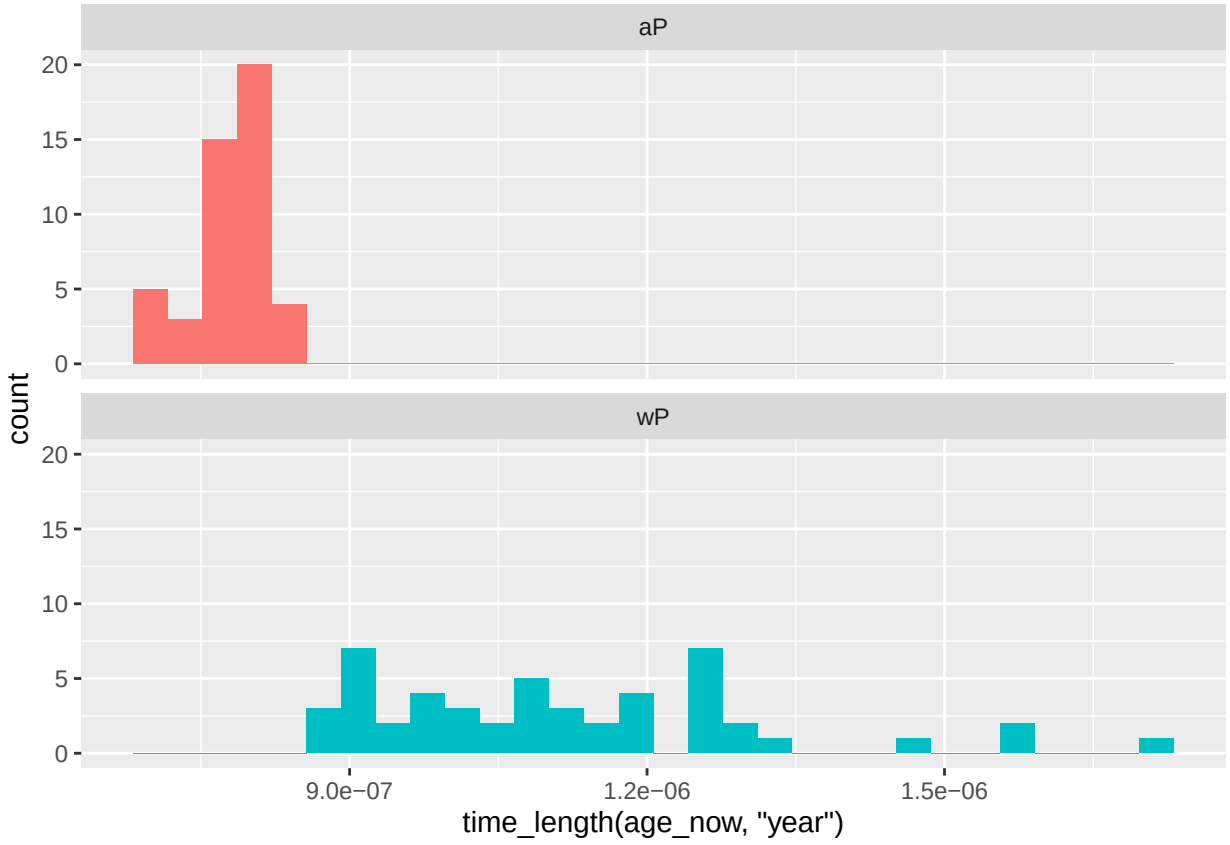
```
##  [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
##  [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
## [17] 36.69815 19.65777 22.73511 32.26557 25.90007 23.90144 25.90007 28.91992
## [25] 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058 24.15058
```

```
## [33] 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876 26.20671
## [41] 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375 22.41752
## [49] 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707 35.65777
## [57] 33.65914 31.65777 25.73580 24.70089 28.70089 33.73580 19.73443 34.73511
## [65] 19.73443 28.73648 27.73443 19.81109 26.77344 33.81246 25.77413 19.81109
## [73] 18.85010 19.81109 31.81109 22.81177 31.84942 19.84942 18.85010 18.85010
## [81] 19.90691 18.85010 20.90897 19.04449 20.04381 19.90691 19.90691 19.00616
## [89] 19.00616 20.04381 20.04381 20.07940 21.08145 20.07940 20.07940 20.07940
```

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

```
ggplot(subject) +
  aes(time_length(age_now, "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`.
```




```
# Or use wilcox.test()
x <- t.test(ap_age_now,wp_age_now)

x$p.value
```

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

```
library(dplyr)
dim(subject)
```

```
## [1] 96  8
```

```
meta <- inner_join(specimen, subject)
```

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

```
dim(meta)
```

```
## [1] 729  13
```

```
View(meta)
```

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    19
```

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
```

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
```

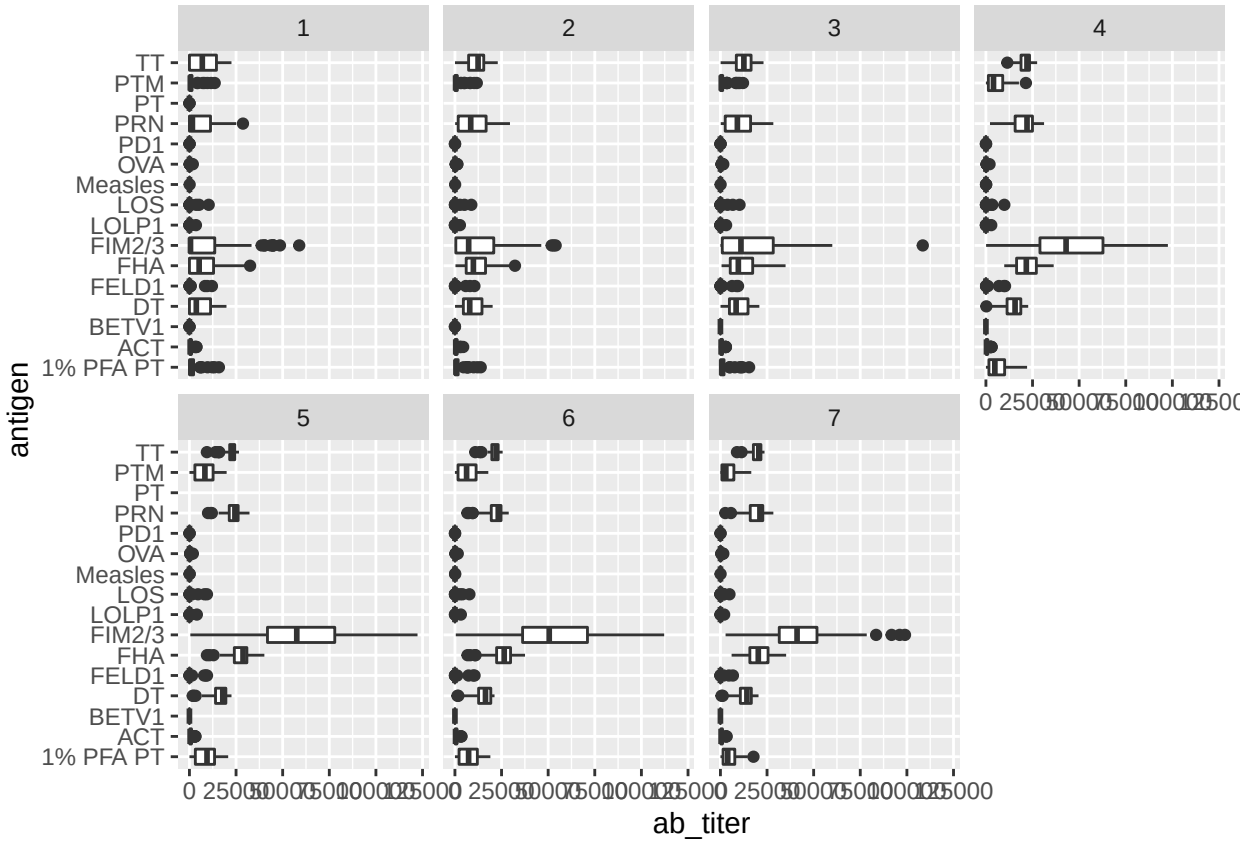
Examine IgG1 Ab titer levels

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

```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1    IgG1                TRUE     ACT 274.355068 IU/ML
## 2           1    IgG1                TRUE     LOS  10.974026 IU/ML
## 3           1    IgG1                TRUE   FELD1   1.448796 IU/ML
## 4           1    IgG1                TRUE   BETV1   0.100000 IU/ML
## 5           1    IgG1                TRUE   LOLP1   0.100000 IU/ML
## 6           1    IgG1                TRUE Measles  36.277417 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                 3.848750          1                           -3
## 2                 4.357917          1                           -3
## 3                 2.699944          1                           -3
## 4                 1.734784          1                           -3
## 5                 2.550606          1                           -3
## 6                 4.438966          1                           -3
##   planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
## 1                             0         Blood     1          wP         Female
## 2                             0         Blood     1          wP         Female
## 3                             0         Blood     1          wP         Female
## 4                             0         Blood     1          wP         Female
## 5                             0         Blood     1          wP         Female
## 6                             0         Blood     1          wP         Female
##                ethnicity  race year_of_birth date_of_boost   study_name
## 1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
```

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

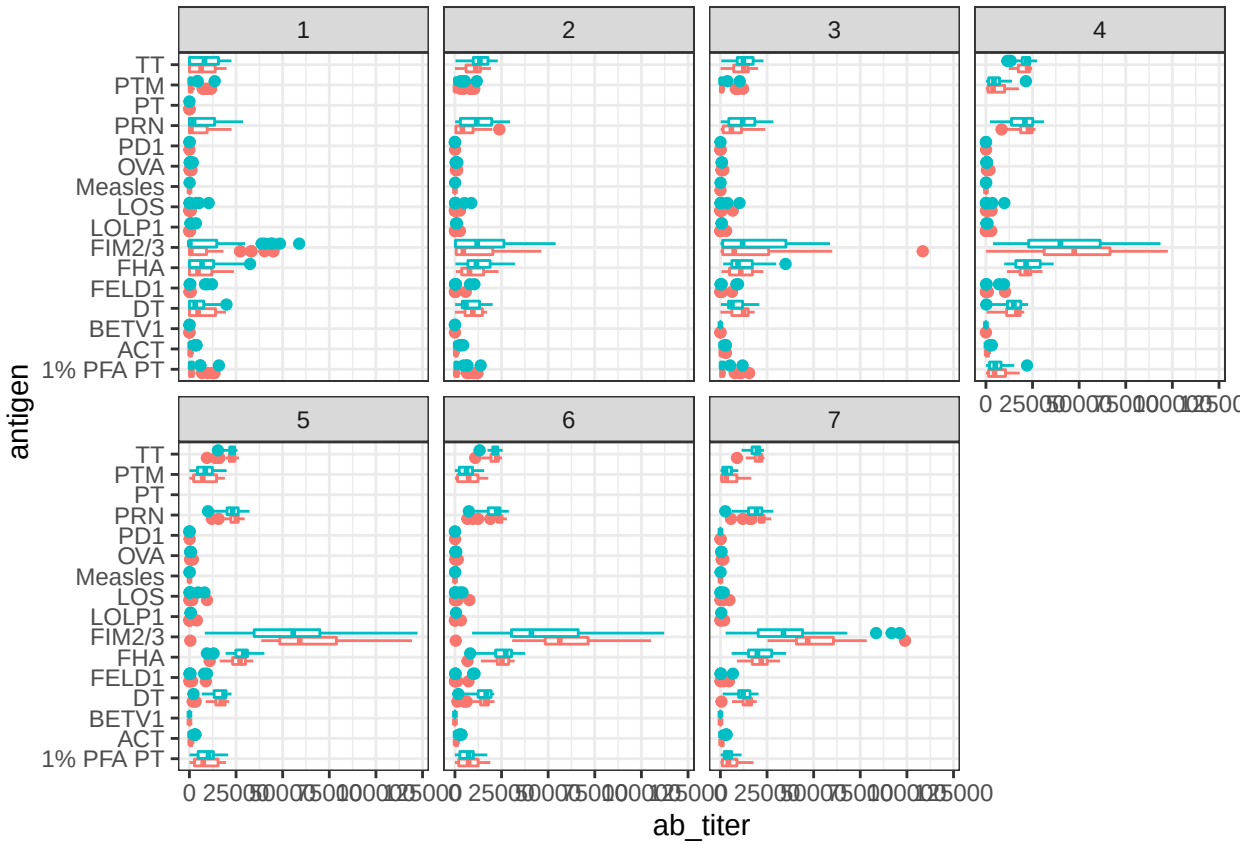
```
ggplot(ig1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



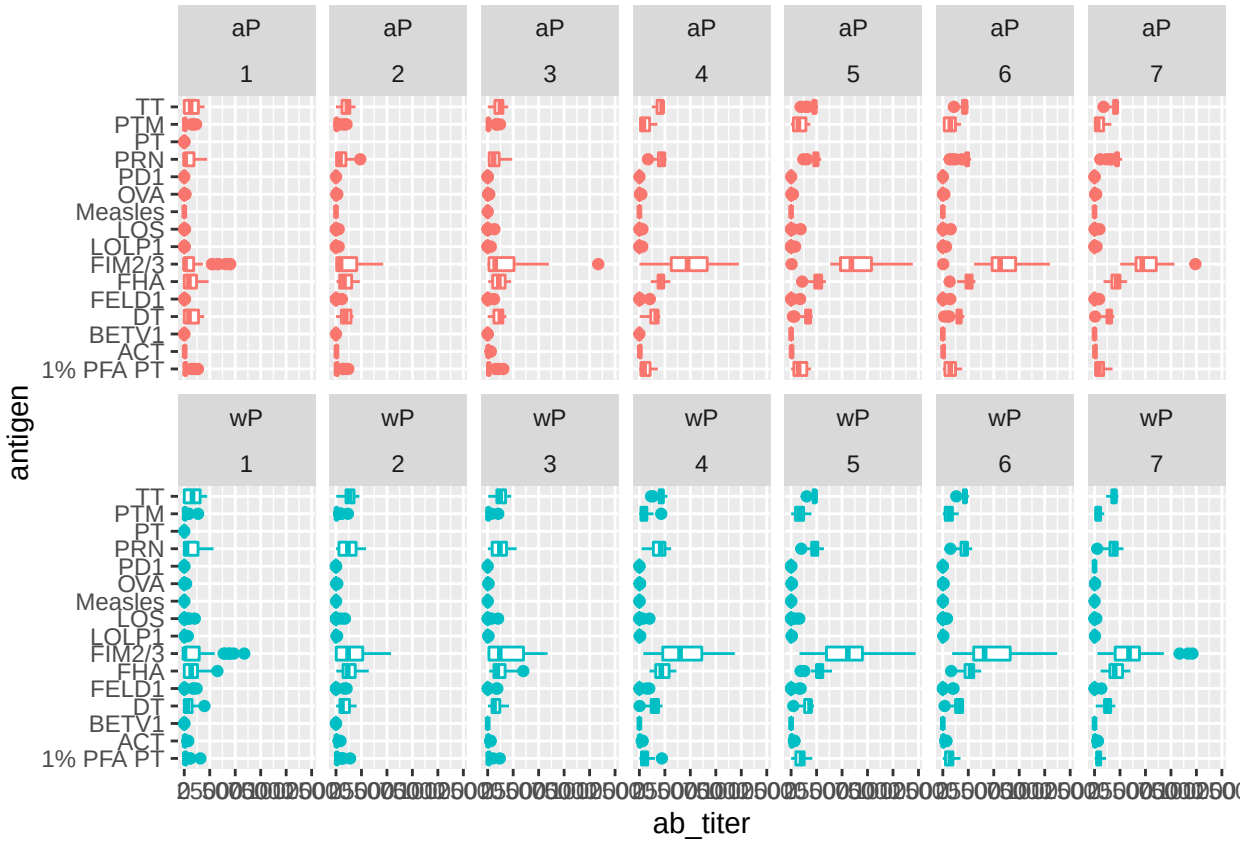
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3,TT and PRN have different levels of antibody but not other. Because FIM2/3 is the antigen the vaccine specificity targeting.

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

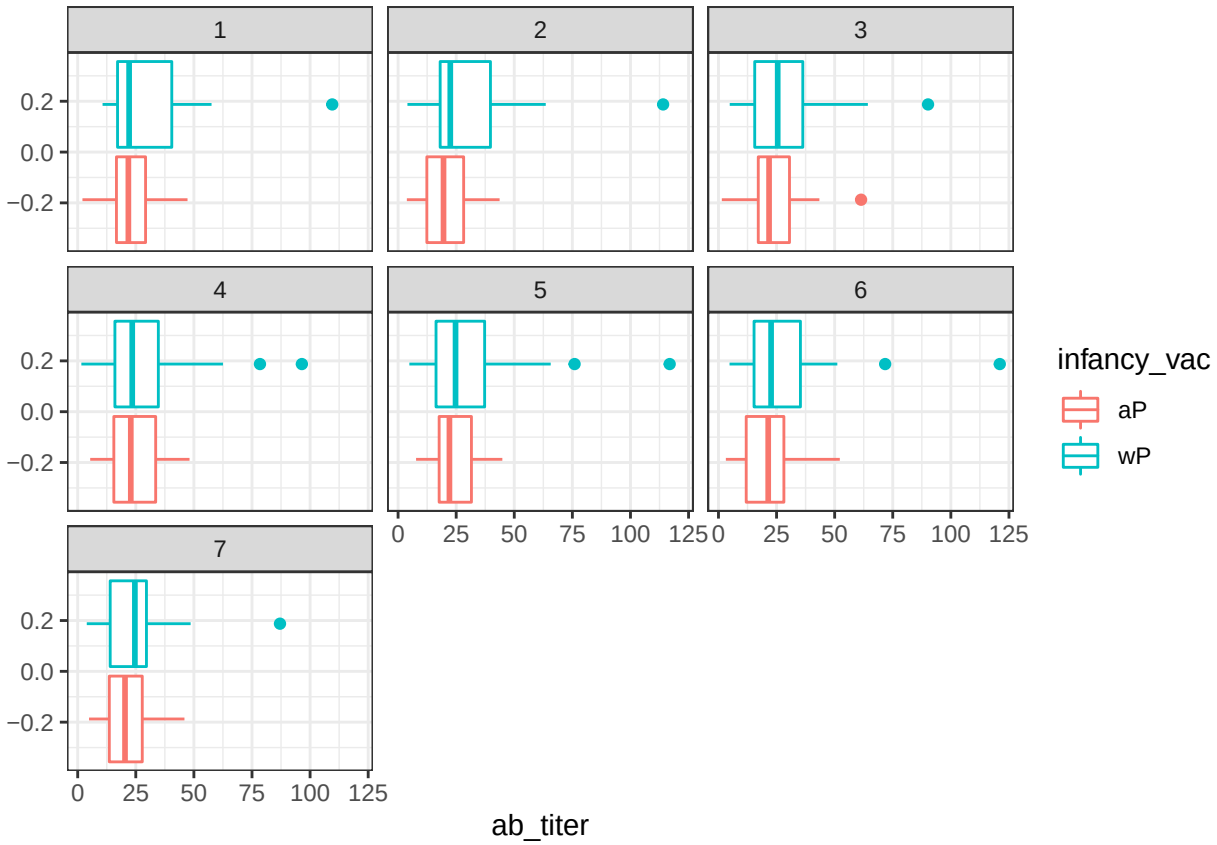



```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



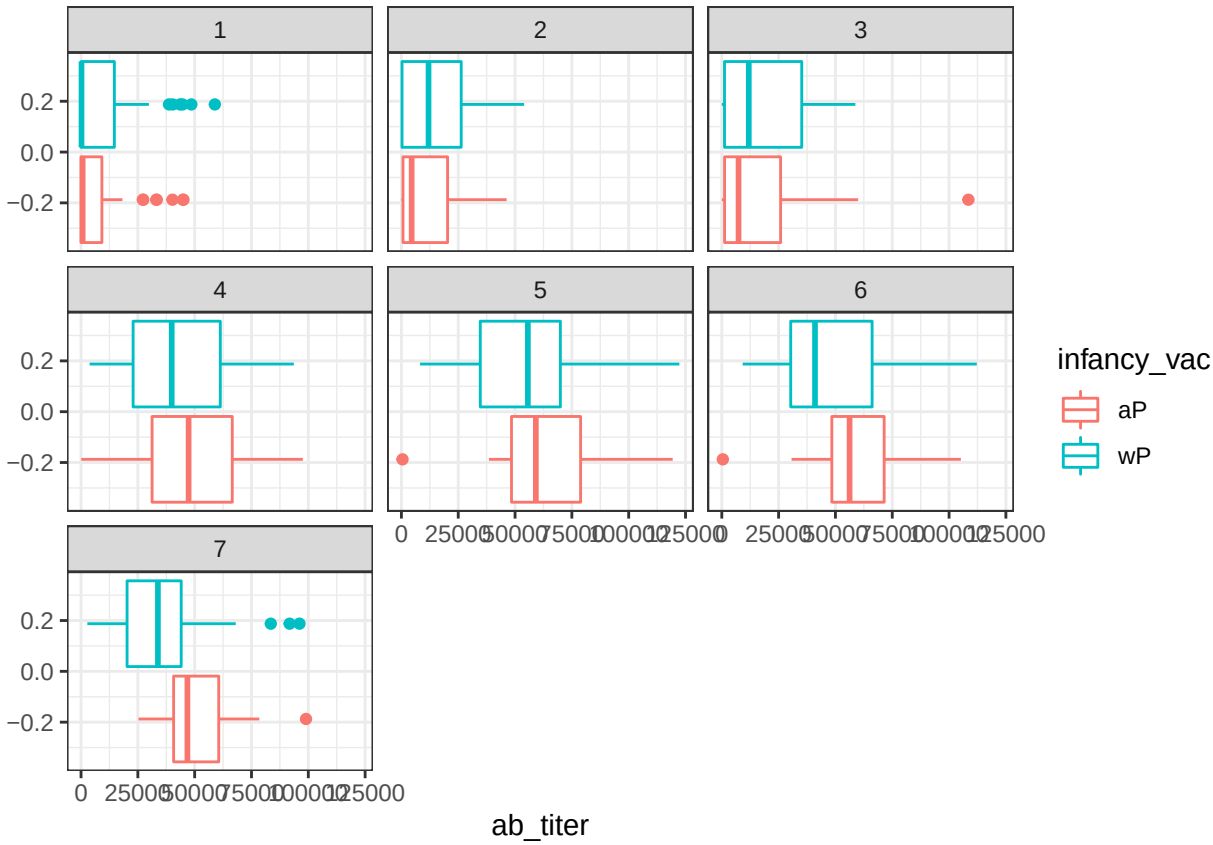
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate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

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



same for antigen=="FIM2/3"

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



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

FIM2/3 has really high time course. Because FIM2/3 is the antigen the vaccine specificity targeting. The vaccine is working.

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

No. it is very little difference in ap vs wp responses. It is hard to draw any conclusions.

Obtaining CMI-PB RNASeq data

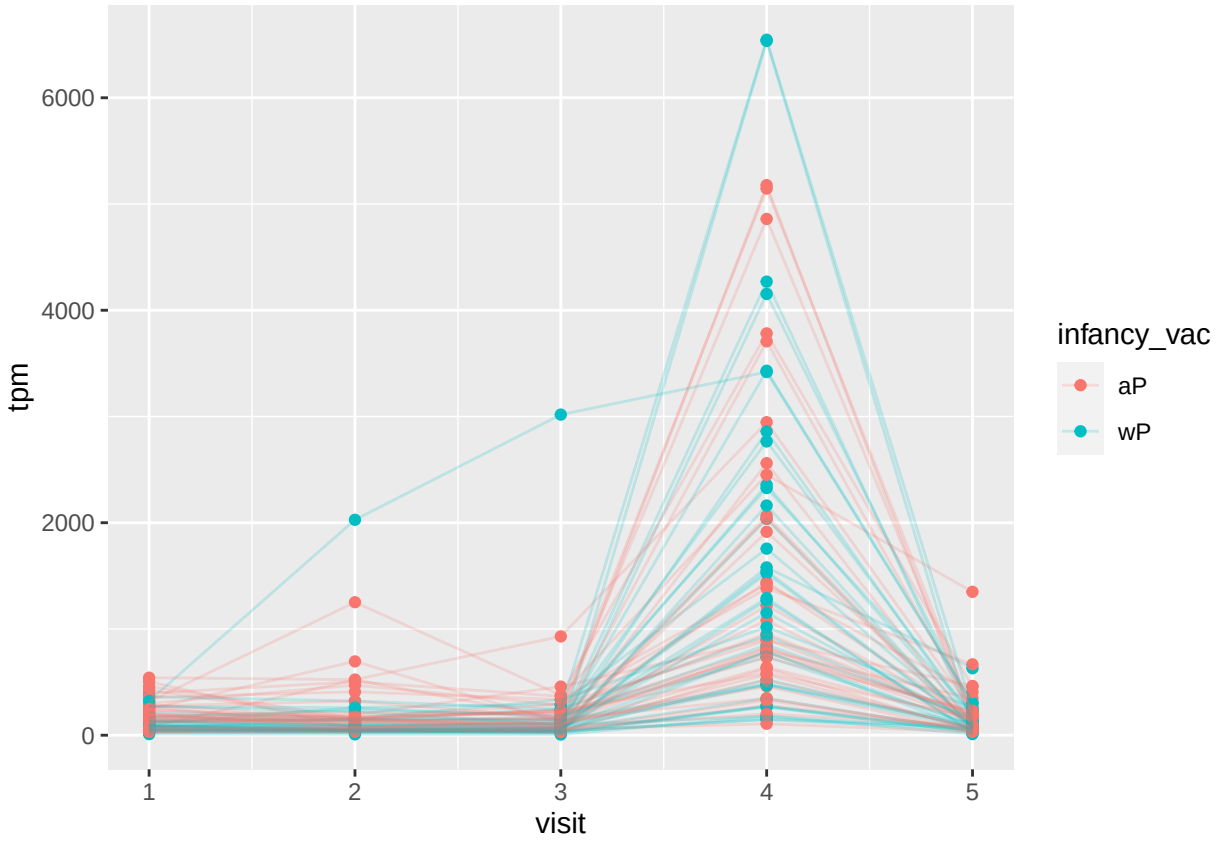
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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

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

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

Q18. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

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



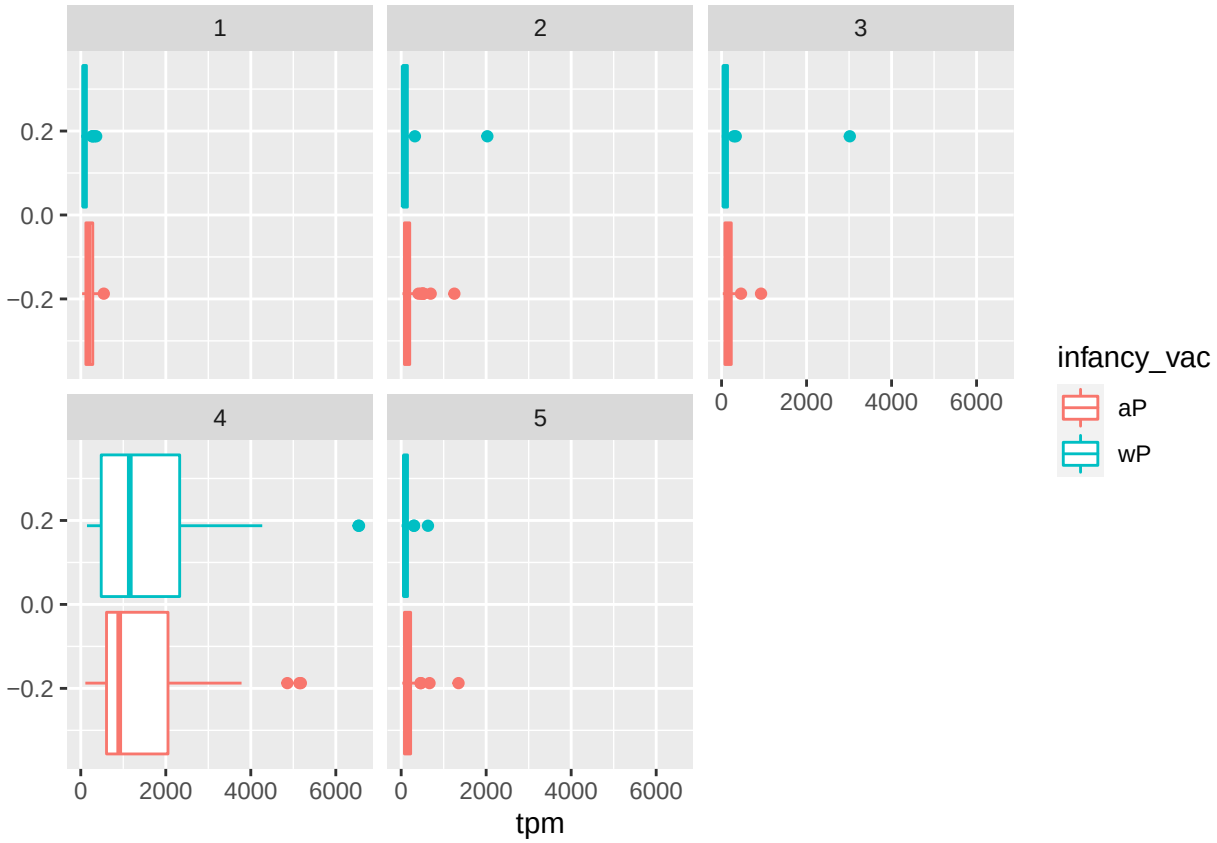
Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

the maximum level is over 6000, at the 4th visit.

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

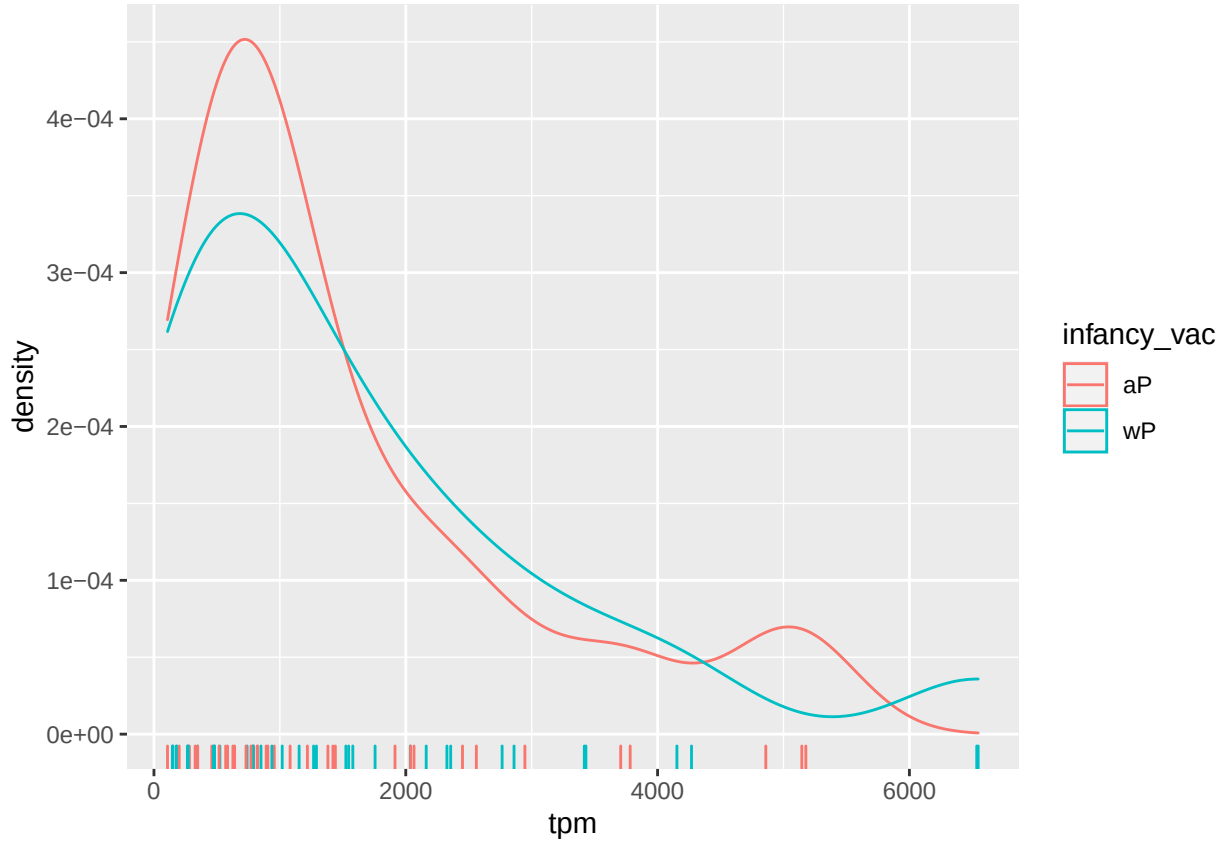
It is not the same pattern. becuase the antibodies can persist inside of the body. antibodies always can detect after a long time. That is the reason RNA level is different from the antibody level. Gene express antibody once a while but antibodies are long persistent inside the body.

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



I do not understand the meaning of the plot. :(

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



Working with larger datasets [OPTIONAL]

```
# Change for your downloaded file path
rnaseq <- read.csv("2020LD_rnaseq.csv")

head(rnaseq,3)
```

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

```
dim(rnaseq)
```

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

```
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
```

```
length(n_genes)
```

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

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

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

```
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
```

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] tidyr_1.2.0     dplyr_1.0.8     lubridate_1.8.0 jsonlite_1.8.0
## [5] ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9        pillar_1.7.0     compiler_4.1.2   tools_4.1.2
##  [5] digest_0.6.29    evaluate_0.15    lifecycle_1.0.1  tibble_3.1.6
```

```
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_1.0.1      cli_3.2.0
## [13] DBI_1.1.2        rstudioapi_0.13  yaml_2.3.5       xfun_0.29
## [17] fastmap_1.1.0    withr_2.4.3      stringr_1.4.0    knitr_1.37
## [21] generics_0.1.2   vctrs_0.3.8      grid_4.1.2       tidyselect_1.1.2
## [25] glue_1.6.2       R6_2.5.1         fansi_1.0.2      rmarkdown_2.11
## [29] farver_2.1.0     purrr_0.3.4      magrittr_2.0.2   scales_1.1.1
## [33] ellipsis_0.3.2   htmltools_0.5.2  assertthat_0.2.1 colorspace_2.0-3
## [37] labeling_0.4.2   utf8_1.2.2       stringi_1.7.6    munsell_0.5.0
## [41] crayon_1.5.0
```