

Class 19: Investigating Pertussis Resurgence

Patrick Tran

1. Investigating pertussis cases by year

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

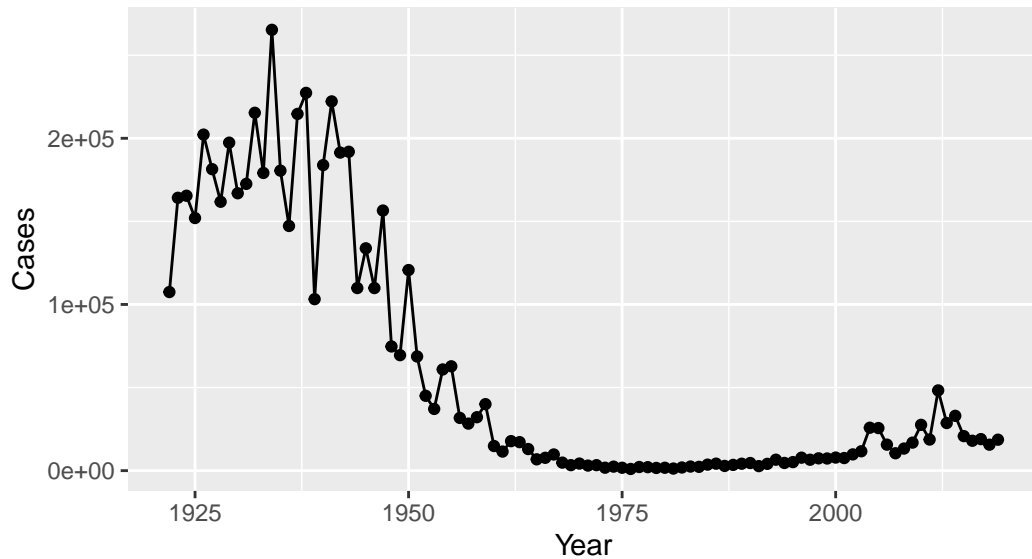
Q1. With the help of the R “addin” package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)
baseplot <- ggplot(cdc) +
  aes(Year,Cases) +
  geom_point() +
  geom_line() +
  labs(title="Cases of Pertussis in US from 1920 to 1999",
        subtitle= "Data from the CDC")

baseplot
```

Cases of Pertussis in US from 1920 to 1999

Data from the CDC



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

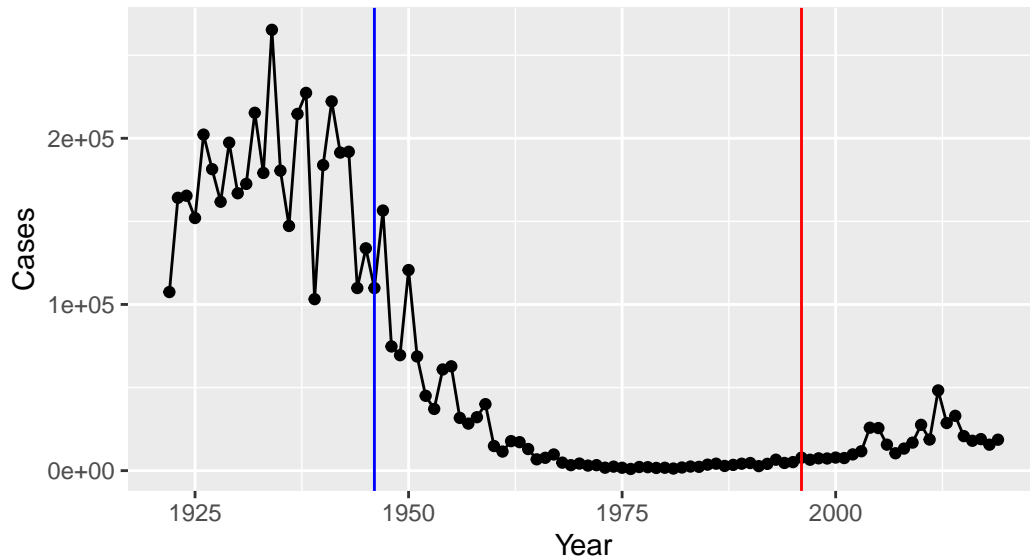
Two types of pertussis vaccines are currently available: whole-cell pertussis (wP) and acellular pertussis (aP).

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, col = "blue") +  
  geom_vline(xintercept = 1996, col = "red")
```

Cases of Pertussis in US from 1920 to 1999

Data from the CDC



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

The number of cases after the introduction of the aP vaccine didn't change instantaneously. However, it can be observed that the cases are increasing after some time. There are spikes or large outbreaks in the cases as well. A possible explanation is bacterial evolution and that vaccine immunity is observed.

3. The CMI-PB Project

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

CMI-PB returns data from it's API in JSON format (like most APIs). We will use the jsonlite package to get data from this API.

```
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	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 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

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)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

First, calculate the age in years of all subjects:

```
age_years <- time_length(today()-ymd(subject$year_of_birth), "years")
subject$age <- age_years
```

Now find the average age of all individuals:

```
mean(subject$age)
```

```
[1] 31.05079
```

Now use dplyr to subset to wP or aP subjects

```
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
mean(ap.age)
```

```
[1] 25.5156
```

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

```
[1] 36.36006
```

T-test

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

Welch Two Sample t-test

```
data: ap.age and wp.age
t = -12.092, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.644857 -9.044045
sample estimates:
mean of x mean of y
 25.51560  36.36006
```

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

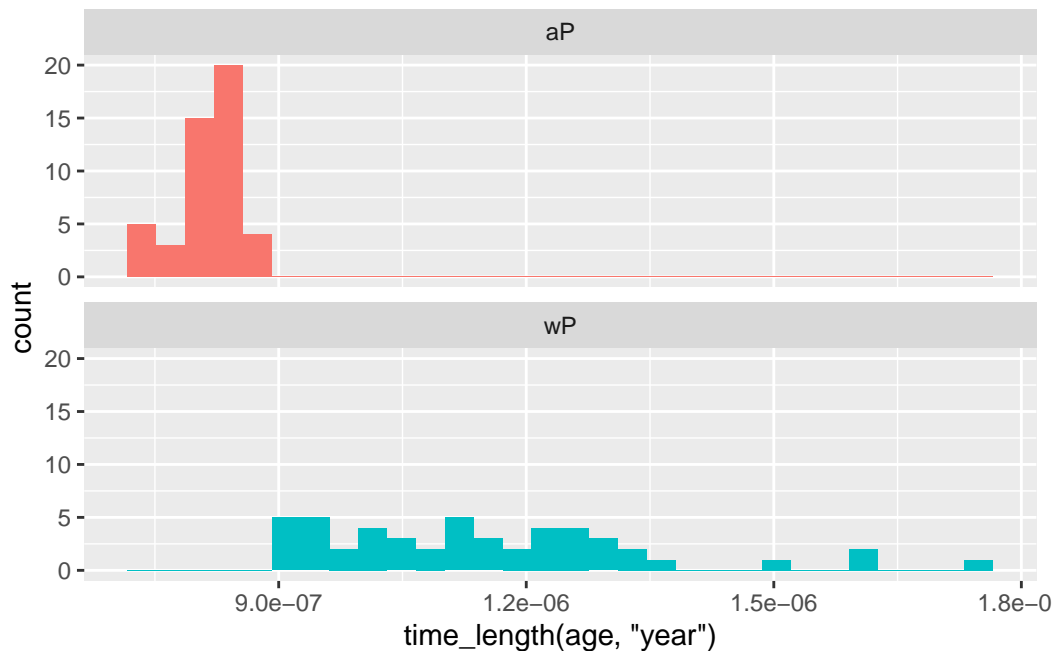
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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`.



Joining multiple tables

Read the specimen and ab_titer tables into R and store data as `specimen` and `titer` named data frames.

```
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 link (a.k.a. “join” or merge) our specimen and subject data frames. The excellent dplyr package (that we have used previously) 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:


```
dim(subject)
```

```
[1] 96  9
```

```
dim(specimen)
```

```
[1] 729  6
```

```
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 37.19644
2 37.19644
3 37.19644
4 37.19644
5 37.19644
6 37.19644

```

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
```

```
head(abdata)
```

```

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 subject_id actual_day_relative_to_boost
1 UG/ML                2.096133           1                -3
2 IU/ML               29.170000           1                -3
3 IU/ML                0.530000           1                -3
4 IU/ML                6.205949           1                -3
5 IU/ML                4.679535           1                -3
6 IU/ML                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	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	37.19644
2	37.19644
3	37.19644
4	37.19644
5	37.19644
6	37.19644

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

The number of visit 8 is very low compared to the other visits. There is not that much data and we should exclude this.

4. 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

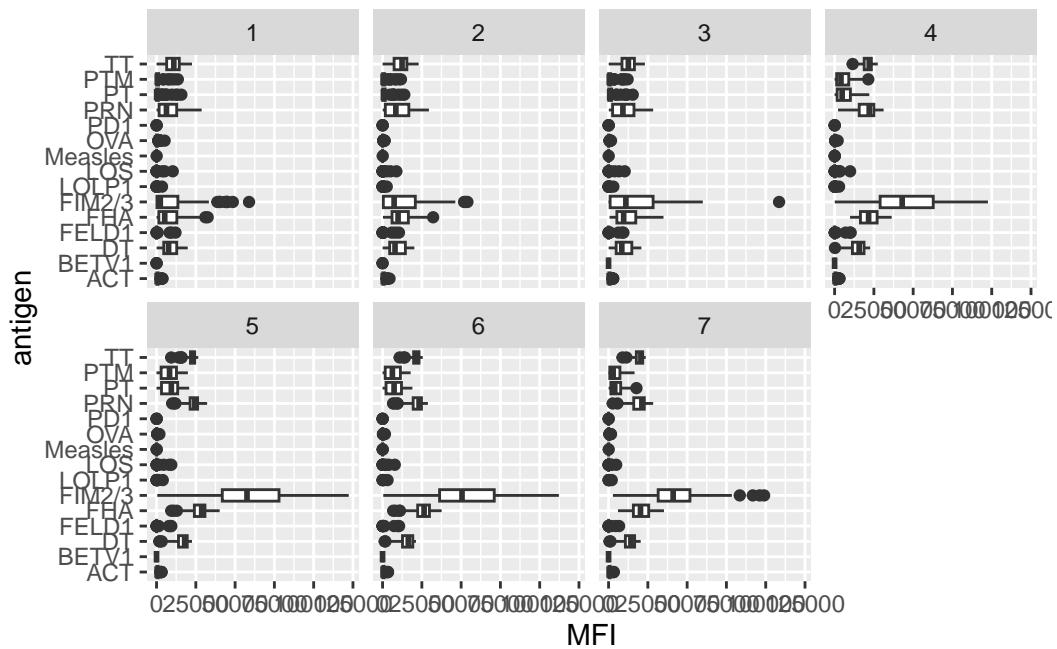
	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Not Hispanic or Latino	White	1986-01-01	2016-09-12	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	37.19644
2	37.19644
3	37.19644
4	37.19644
5	37.19644

6 37.19644

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

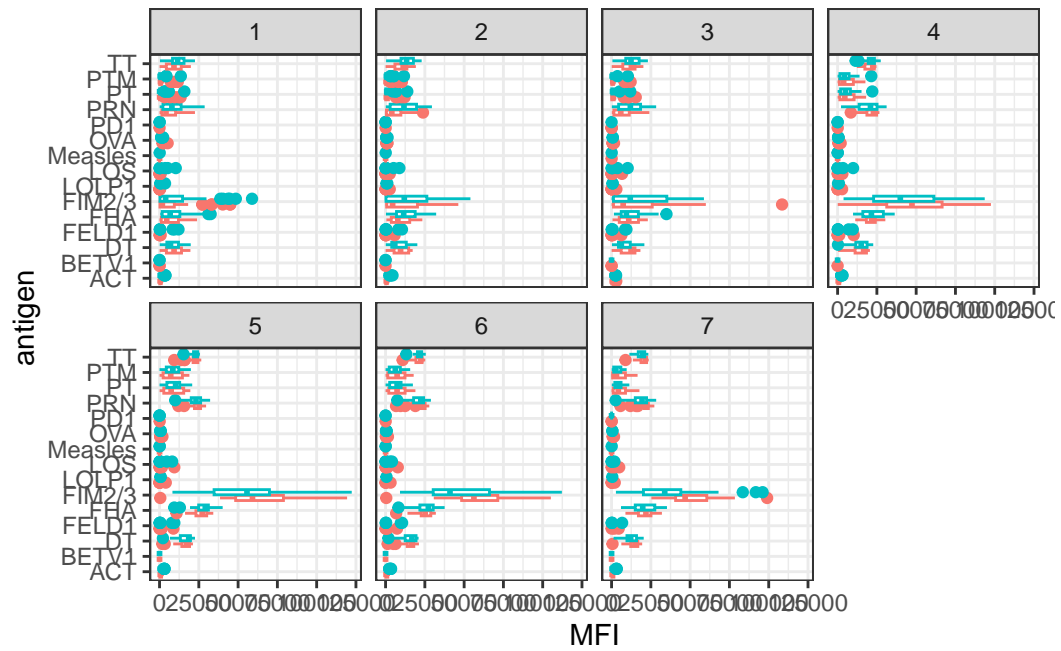
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



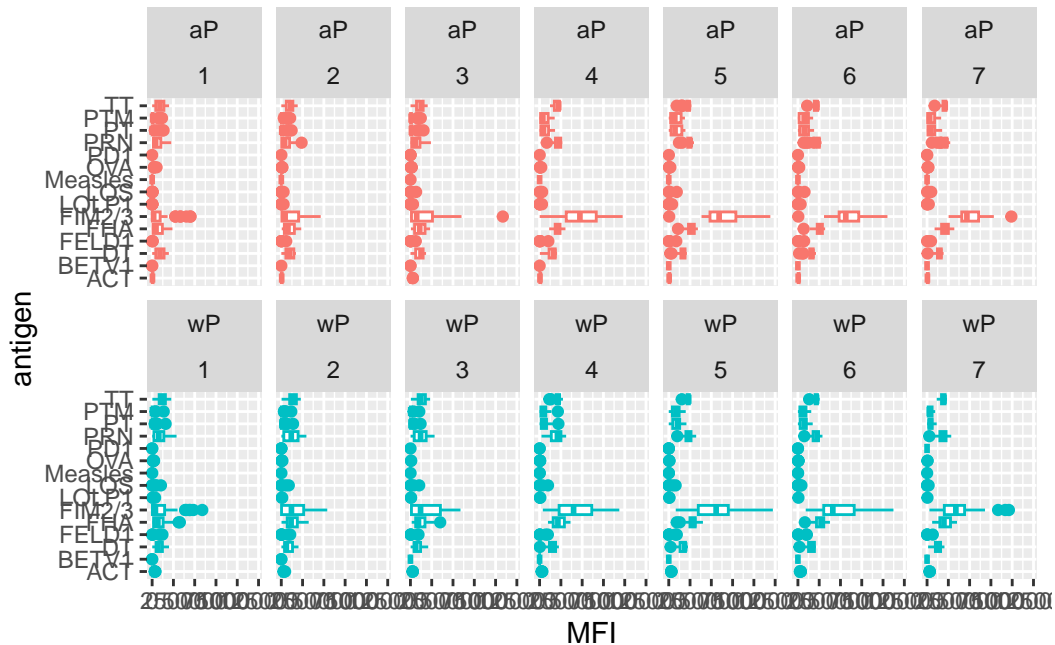
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

The main differences are for FIM2/3, FHA, PT all of these are in the aP boost vaccine.

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



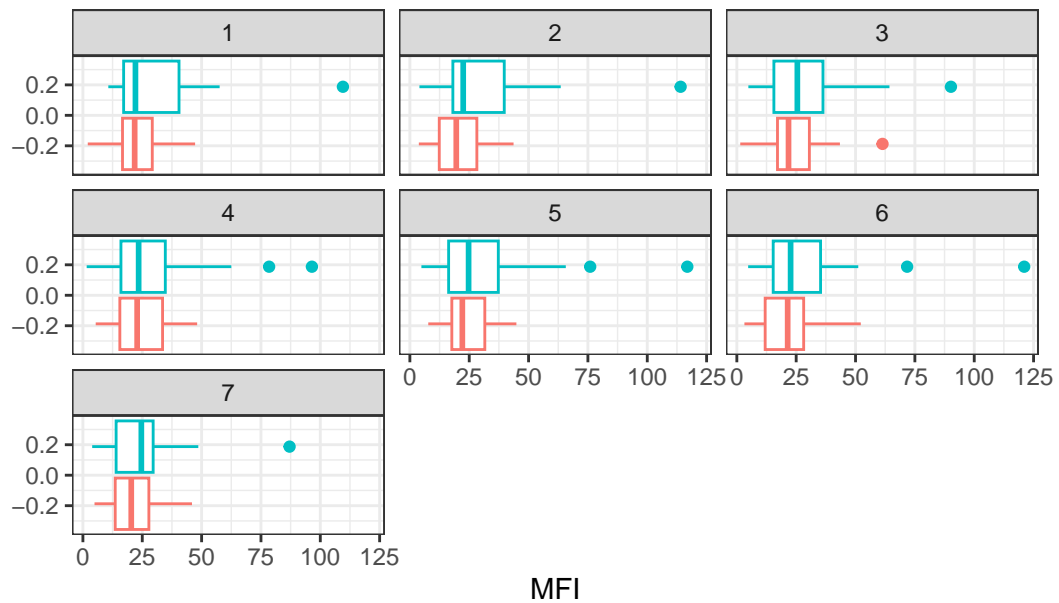
```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



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).

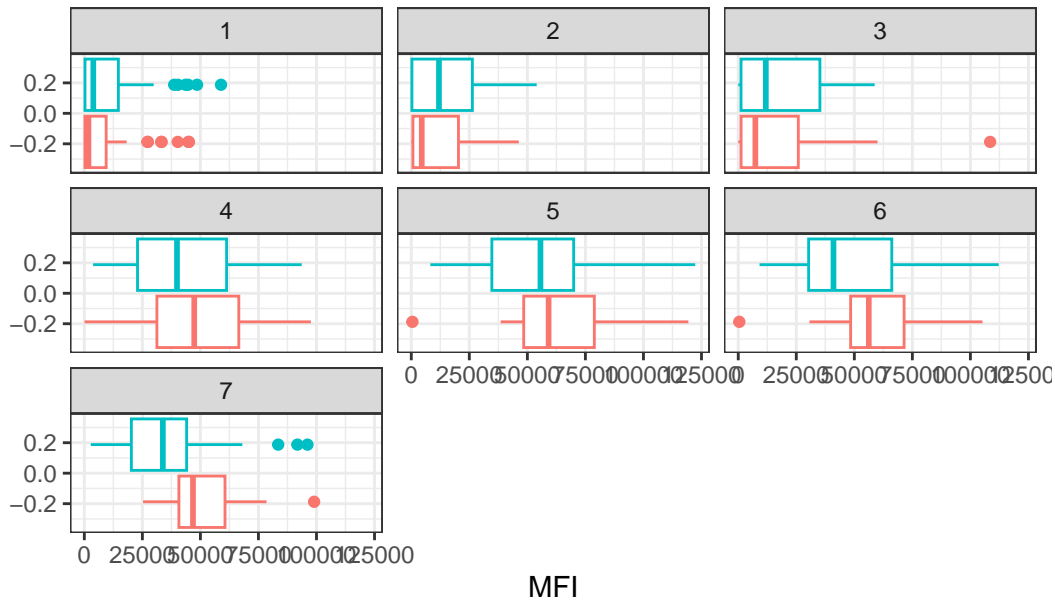
```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  labs(title="Measles antigen levels per visit (aP red, wP teal)") +
  theme_bw()
```

Measles antigen levels per visit (aP red, wP teal)



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


FIM2/3 antigen levels per visit (aP red, wP teal)



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

The FIM2/3 data shows that the levels increase over time. It increases more compared to the Measles data. For both aP and wP subjects, the levels appear to peak at visit 5 and then decrease afterwards.

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

The wP responses in the Measles data was slightly greater than the aP responses for all visits. The FIM2/3 data shows that is wP response is greater until visit 3. After visit 3, the wP response is less compared to the aP response.

5. Obtaining CMI-PB RNASeq data

For RNA-Seq data the API query mechanism quickly hits the web browser interface limit for file size. We can do “targeted” RNA-Seq queries via the web accessible API.

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENS00000211896."
rna <- read_json(url, simplifyVector = TRUE)
```

```
head(rna)
```

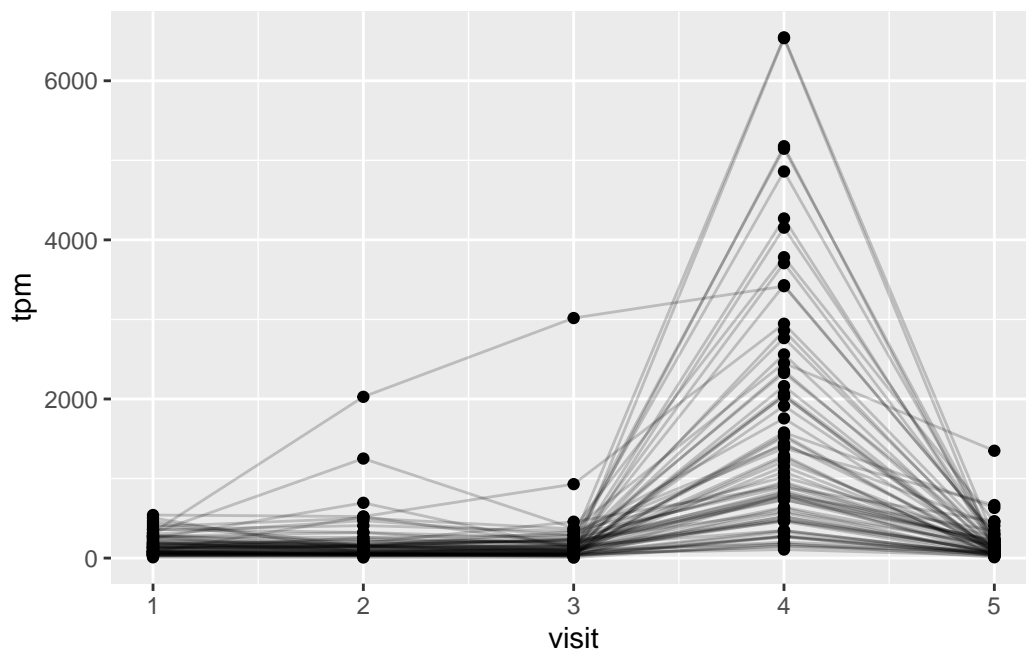
	versioned_ensembl_gene_id	specimen_id	raw_count	tpm
1	ENSG00000211896.7	344	18613	929.640
2	ENSG00000211896.7	243	2011	112.584
3	ENSG00000211896.7	261	2161	124.759
4	ENSG00000211896.7	282	2428	138.292
5	ENSG00000211896.7	345	51963	2946.136
6	ENSG00000211896.7	244	49652	2356.749

```
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)
```



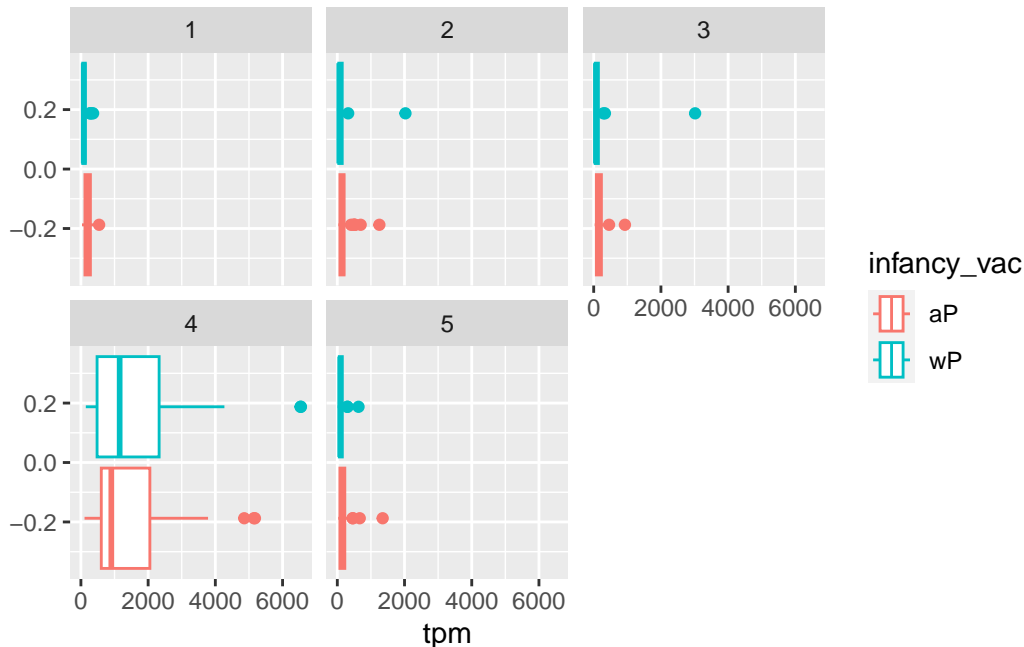
Q19. What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

A pronounced peak is at visit 4. There is a very rapid decline afterwards to low levels of transcription.

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

This does not match with the antibody titer data. The antibodies are long-lived and do not decline at the same rate of this gene expression figure.

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



Focus in on visit 4 and facet by aP/wP subjects

```
ssrna %>%  
  filter(visit==4) %>%  
  ggplot() +
```

```
aes(tpm, col=infancy_vac) + geom_density() +  
geom_rug()
```

