Apparent strong selection in young children as an artifact of longer shedding durations

Mike Famulare, PhD
Institute for Disease Modeling
Gates Foundation
mike.famulare@gatesfoundation.org

2025-05-09

This is a quick simulation experiment to demonstrate a hypothesis for a very interesting observation in Frida Belinky's talk at Dynamics and Evolution of Viruses titled: *Unveiling novel hCoV-OC43 genotypes: the hidden world of intra-host evolution.*

The question: The observation that caught my attention, as I understood it, was that there were higher variant frequencies at a particular antigenic site among young children than among older children and adults. This is superficially counterintuitive because young children have little if any prior immunity to OC43, whereas older people very often have prior adaptive immunity from prior infections. Thus one might expect apparent immune escape selection to be to be stronger in older ages, which appears to be counter to the data.

Hypothesis: If one assumes observed high variant frequencies at an antigenic site are primarily due to immune selection, then this would be a surprising result. However, there is a sampling bias phenomenon that can explain this observation, and it works as follows. Young children who have had zero (or few) prior infections shed virus for longer periods of time on average. Any analysis that analyzes only virus-positive cases will tend to sample later in infections for people who shed longer. Thus, children will be more likely to show higher variant frequencies, even if there is no within-host selection variation with age, *other than the overall impact of prior immunity on infection duration.*

This simulation: To demonstrate what the OC43 variant frequency vs age could look like under the assumption that the effect is driven entirely by the correlation of shedding duration with age, without any difference in the strength of selection by age, I mocked up this simulation. How it works is as follows:

- First, I define a cohort of individuals with randomly sampled shedding durations, from a distribution that looks roughly like what I recall is appropriate for OC43 (but I didn't look up any references).
- Second, for each individual, I uniformly sampled a sampling time that represents when they got picked up during their infection and tested positive. While the sampling time is random during an infection, sampling will on average be later during infections for longer shedding durations. This is the key feature of cross-sectional sampling for positive infections that are relevant to this model of the variant frequency phenomenon.
- Third, I define a Wright-Fisher model for a single amino acid. I picked parameters that are coronavirus-like
  - Ne 100,
  - $\mu = 8.5e-4$ per site per year,
  - a 12 hour within-host generation time and
  - an assumed selection coefficient of $s = 0.1$, chosen to look about right and is fixed, independent of any other variable (thus assuming no effect of individual immunity on selection).

- Fourth, I made up a simple model of age correlated with shedding duration, such that younger people tend to have longer shedding durations.
- Fifth, I simulated the Wright-Fisher effective population size as a crude proxy for variation in viral load, and correlated this with infection duration, as is typical for viruses.

Results: The main result is how the observed variant frequency depends strongly on when the sample is taken during infection but much more weakly on the viral load. Insofar as longer shedding durations and thus later sampling times are more common among younger ages, we see higher variant frequencies at younger ages. This is qualitatively consistent with the observation reported in the talk yesterday, and does not require a within-host differential immune selection mechanism. While there may also be selection differences by age, that needs direct evidence beyond the age-variant association to demonstrate it.

Setting up the simulation

To start, let's set up our environment

```
library(tidyverse)
library(ggExtra)

set.seed(100) # for reproducibility. comment out to see variation.
```

and define the parameters of the simulation. We start with the two independent components, shedding duration and the baseline Wright-Fisher model.

```
# lognormal for shedding duration
log_shedding_duration_median=log(8) # log days
log_shedding_duration_sd = log(1.3) # log days such that most between 4-15 days

# single amino acid wright fisher
Ne_median = 100 # median intrahost effective pop size
mu = 8.5e-4 # mutation rate per site per day
s = 0.1  # assumed selection coefficient within host
tau = 1/2 # 12 hour generation time in days
```

Then, we start to build up the model of attributes that correlate with shedding duration. Here is a simple model of how viral load, as represented by the effective population size of the Wright-Fisher model (Ne), might be correlated with shedding duration. This roughly reproduces the observation that longer shedding durations tend to be weakly correlated with viral load across all data on coronavirus (although not necessarily in any one study with often small-N). In real biology, this relationship is mediated by a mix of random chance/host factors and prior immunity, and is further overdispersed by the weak correlation between qPCR and reproductive effective population size.

```
# Ne, duration correlation model
rho_Ne_dur=0.7   # guess that feels about right given data I've seen across pathogens over the years
log_sd_Ne_dur = 0.4  # let Ne vary from 50 to 500ish around median 100
```

Finally, we set up the model of correlation between age and shedding duration. This is just cooked up to show the phenomenon, but isn't particularly realistic about how narrow the age of first infection should be on average.

```
# age, duration correlation model
rho_age_dur = -0.5   # negative correlation between age and shedding duration
range_age_dur = 30 # years
```

Running the model

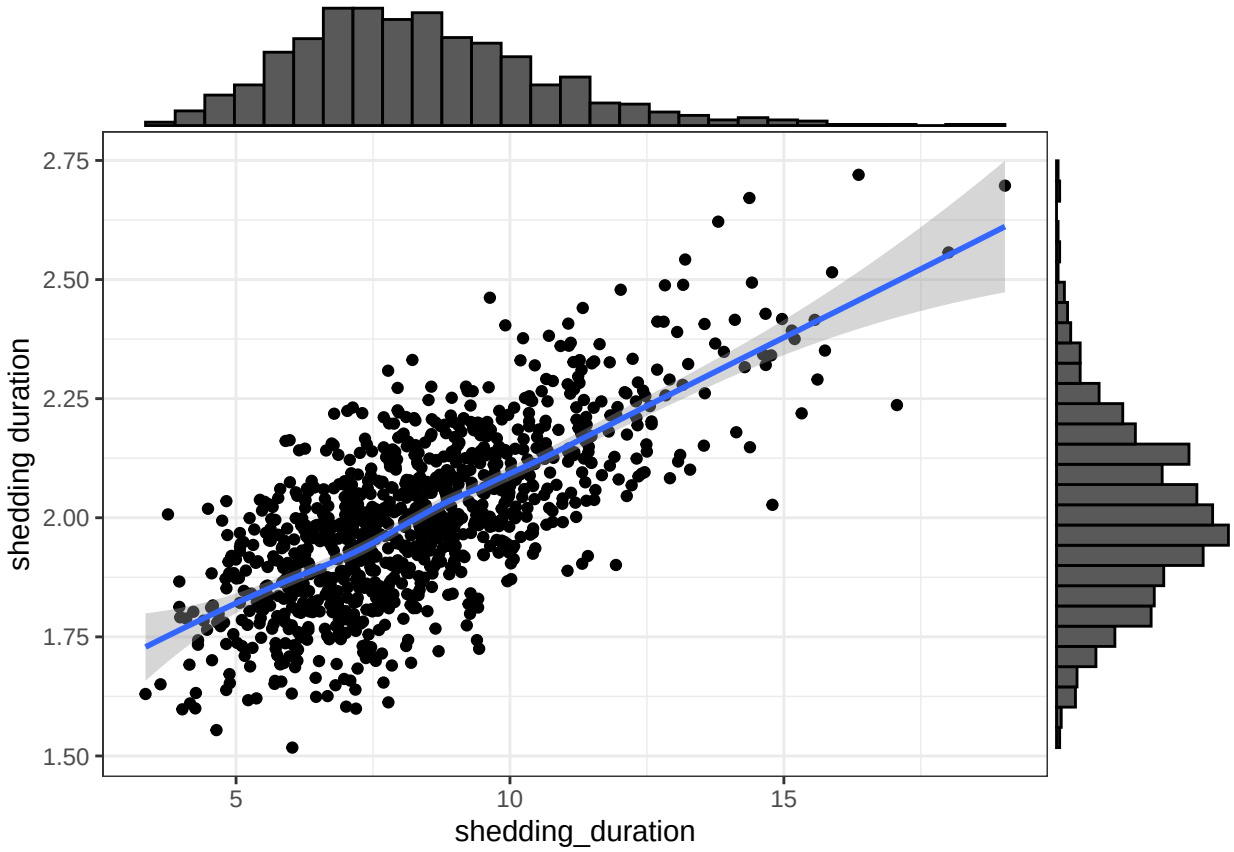
Now that we've set up the population, their (unobserved) shedding durations, (unobserved) sampling time relative to the start of their infection, and observed Ne and ages, we can run the simulation and sample variant frequencies at the amino acid site of interest for each person.

```
# set up the population
N_people=1000
sim_data = data.frame(id=1:N_people,
                      shedding_duration =
                        exp(rnorm(n=N_people,
                                  mean=log_shedding_duration_median,
                                  sd=log_shedding_duration_sd))) |>

  # sample randomly during shedding duration to represent
  # randomly finding positive people during their infections
  mutate(sampling_time = runif(N_people)*shedding_duration) |>

  # sample Ne (proxy for viral load) which is assumed to be correlated with shedding duration
  mutate(Ne = exp(log(Ne_median) + log_sd_Ne_dur*rho_Ne_dur*scale(shedding_duration) +
                    log_sd_Ne_dur*sqrt(1-rho_Ne_dur^2)*rnorm(N_people))) |>

  # sample ages correlated with shedding duration
  mutate(age =  range_age_dur*rho_age_dur*pnorm(scale(shedding_duration)) +
           range_age_dur*sqrt(1-rho_age_dur^2)*exp(runif(nrow(sim_data)))) |>
  mutate(age = age - min(age)) # shift age to be >0 because this is a crappy model
```

Here's what these attributes look like. The correlations are shown by the scatter plots, and the marginal distributions are on the relevant axes. These are chosen to look coronavirus-like, but could be fit to data and better referenced to the literature to be more realistic.

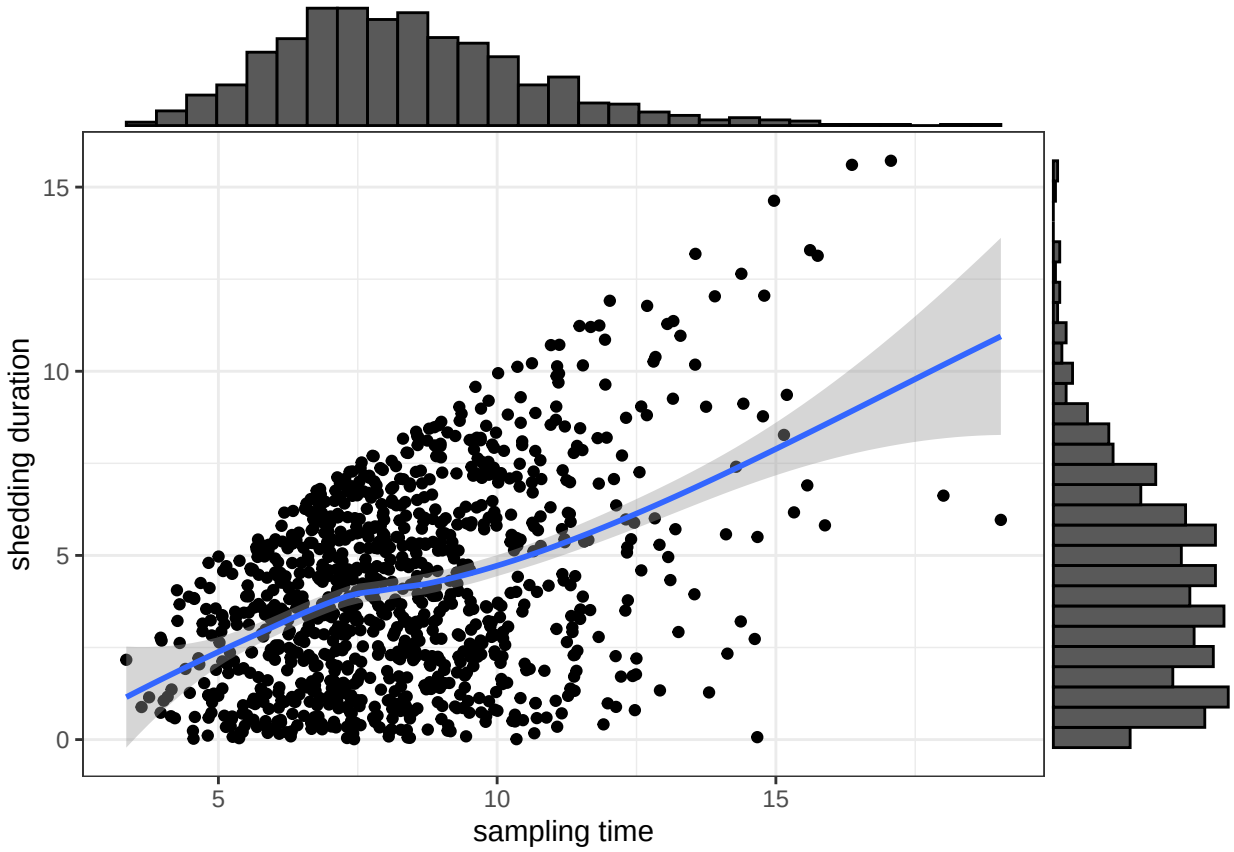
First, we have the unobserved association between shedding duration and effective population size (a proxy for viral load).

```
p =ggplot(sim_data,aes(x=shedding_duration,y=log10(Ne))) +
  geom_point() + geom_smooth(method='loess') + theme_bw()  +
  ylab('shedding duration')
ggMarginal(p, type = 'histogram')
```



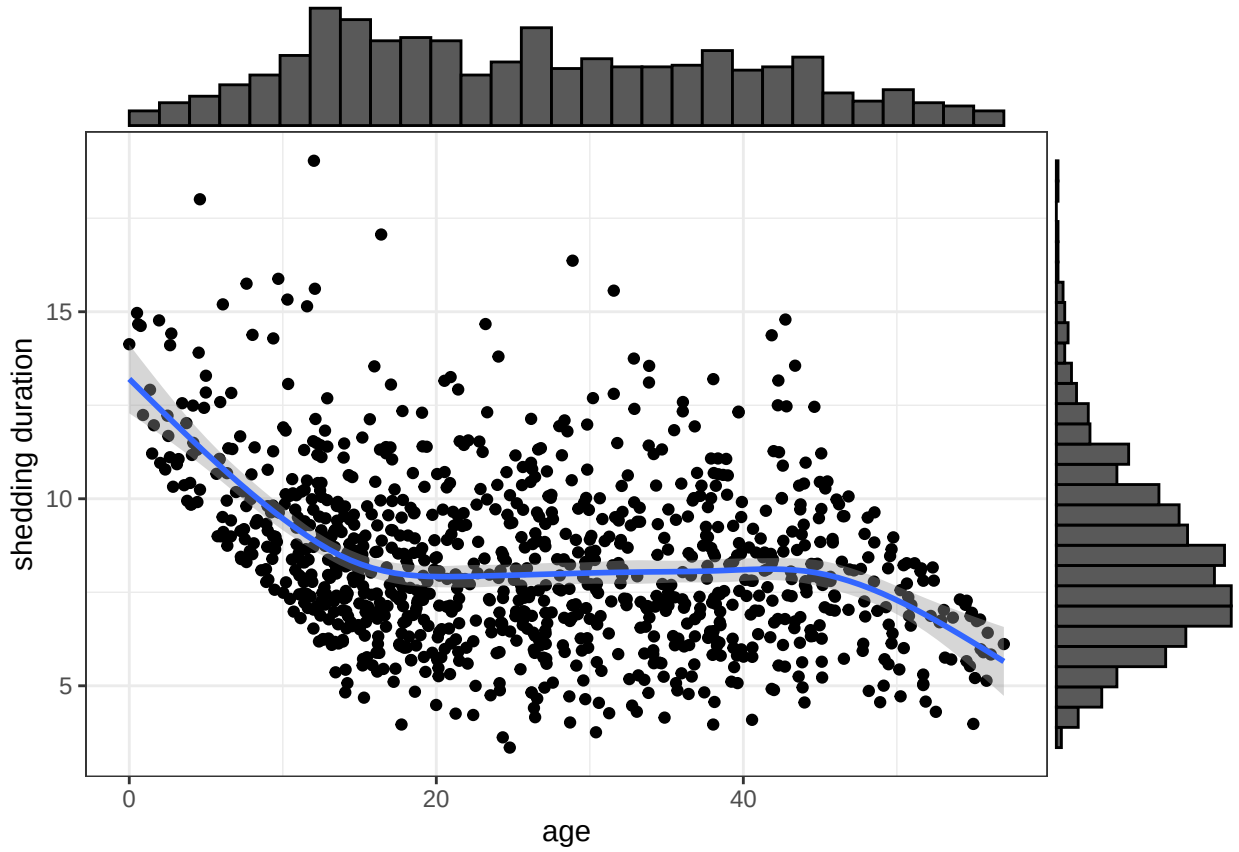
Then, the unobserved association between sampling time and shedding duration. This shows the assumption that when randomly sampling people and finding them positive, sampling times will on average be later in the infection when shedding duration is longer. This is simply survival bias.

```
p =ggplot(sim_data,aes(x=shedding_duration,y=sampling_time)) +
  geom_point() + geom_smooth(method='loess') + theme_bw() +
  ylab('shedding duration') + xlab('sampling time')
ggMarginal(p, type = 'histogram')
```



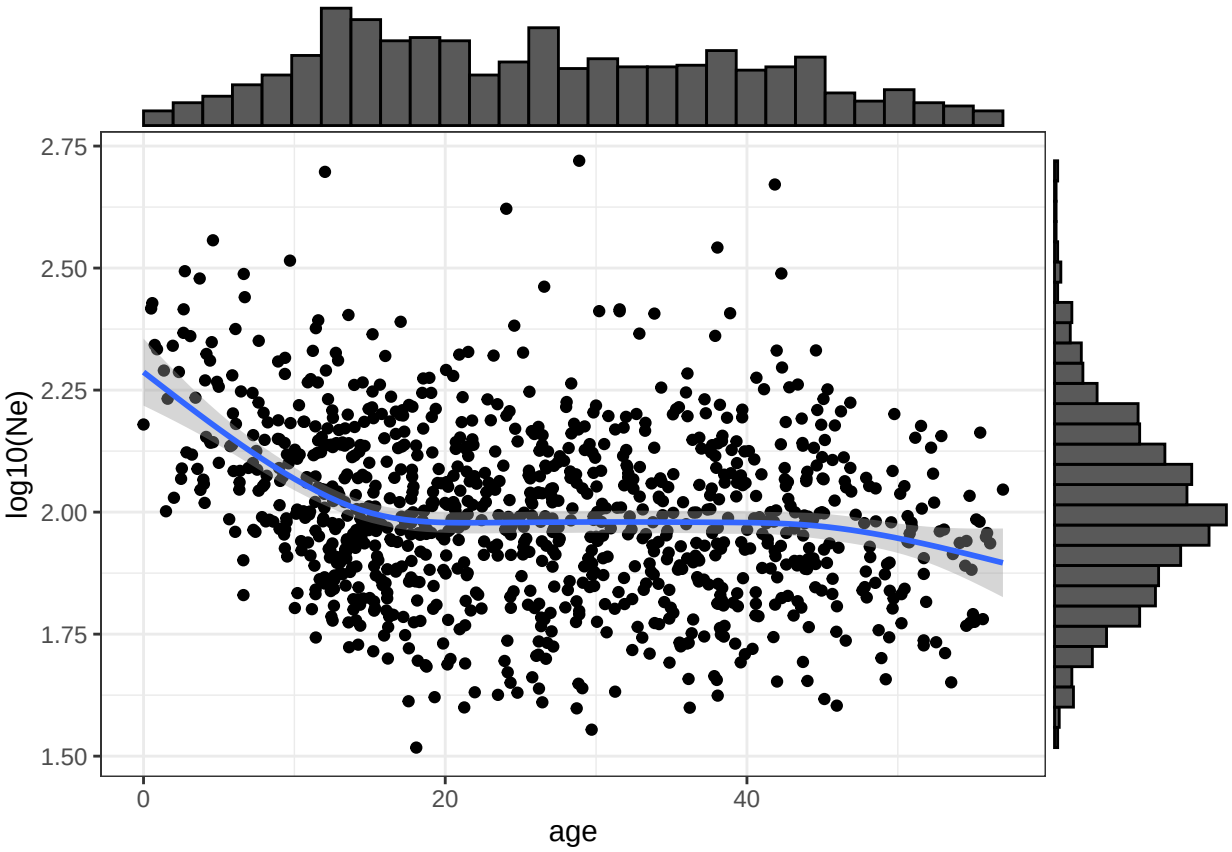
Third is the assumed correlation of shedding duration with age. This is a simple model of how shedding duration tends to be longer when prior immunity is lower. This is not observed in a cross-sectional single sample, but is known to be nearly universal among viral pathogens when individuals are followed longitudinally or following known exposures in contact tracing.

```
p=ggplot(sim_data,aes(y=shedding_duration,x=age)) +
  geom_point() + geom_smooth() + theme_bw()  +
  ylab('shedding duration')
ggMarginal(p, type = 'histogram')
```



And finally, the only observable relationship is age vs viral load. This is a much weaker correlation, as is typical in observational studies.

```
p=ggplot(sim_data,aes(x=age,y=log10(Ne))) +
  geom_point() +
  geom_smooth() + theme_bw()
ggMarginal(p, type = 'histogram')
```



Last, let's run the model to sample people and their observed variant freqencies at the time of sampling.

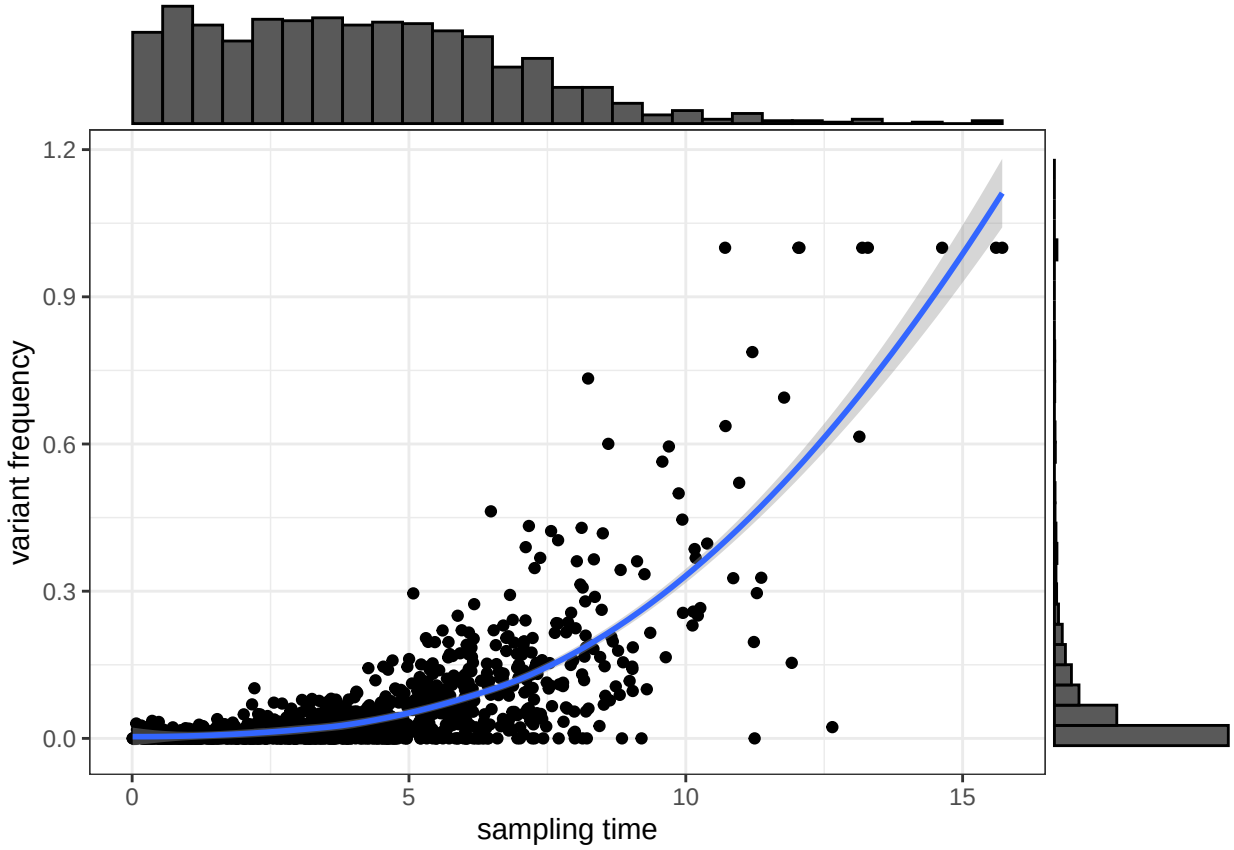
```
## simulate variant fraction at the sampling time and at the amino acid of interest for each person
for (k in 1:N_people){
  tmp_var_count=0
  for (n in 1:floor(sim_data$sampling_time[k]/tau)){
    tmp_var_count =
      # previous generation count of variants grows better
      tmp_var_count*exp(s) +
      # new variants from remaining orginal allele
      exp(s)*rpois(1,lambda = mu*(sim_data$Ne[k]-tmp_var_count)/tau) +
      # reversion
      -exp(-s)*rpois(1,lambda = mu*tmp_var_count/tau)

    # select but keep within Ne
    tmp_var_count=min(sim_data$Ne[k],tmp_var_count*(1+s))
  }
  sim_data$var_freq[k]=tmp_var_count/sim_data$Ne[k]
}
```

Results

By design, the strongest correlation is with the sampling time during an infection. The later in the infection the virus is sampled, the higher the variant frequency on average. This isn't interesting, but it is a demonstration that the model does what it's supposed to.

```
p=ggplot(sim_data,aes(x=sampling_time,y=var_freq)) +
  geom_point() +
```

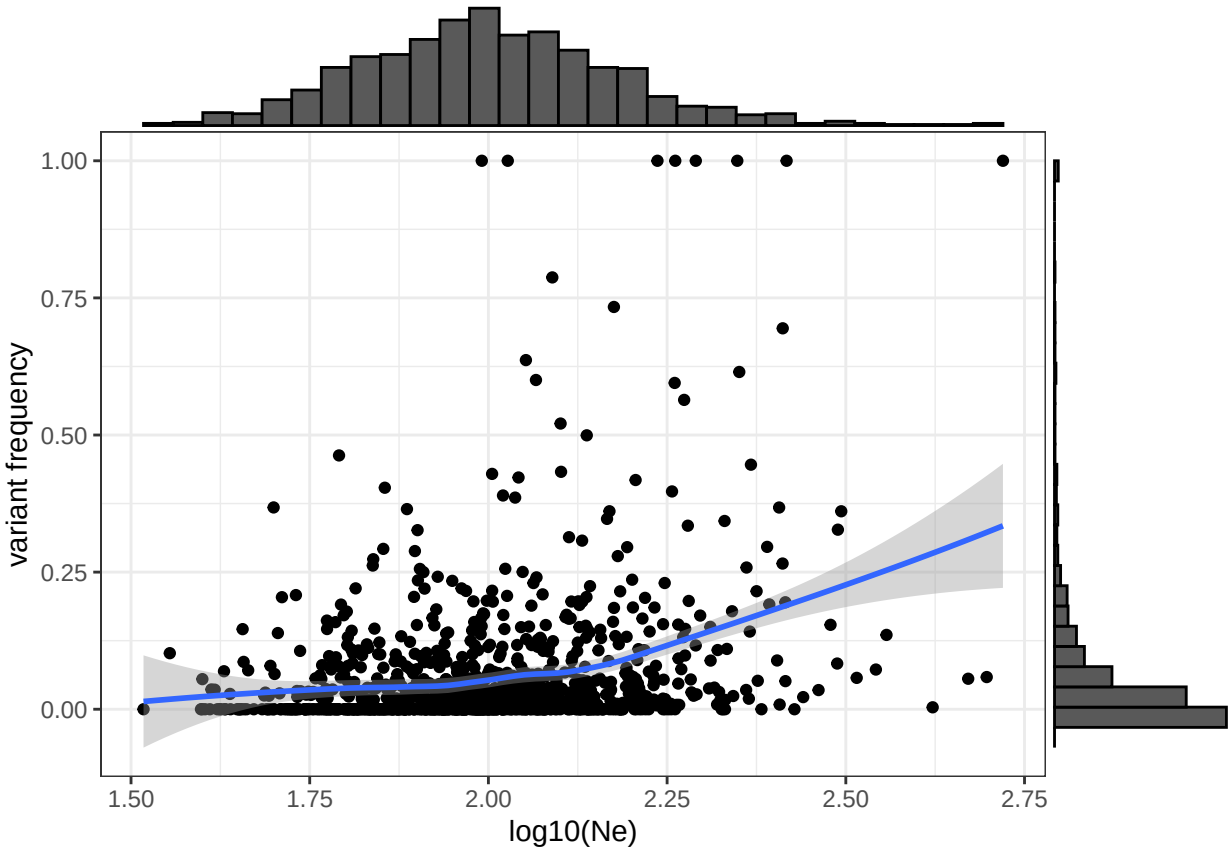
```
  geom_smooth(method='loess') +
  # geom_smooth(method='lm',color='red') +
  theme_bw() +
  ylab('variant frequency') + xlab('sampling time')
ggMarginal(p, type = 'histogram')
```




```
# #+ shedding_duration_vs_var_freq, echo=TRUE, message=FALSE
# p=ggplot(sim_data,aes(x=shedding_duration,y=var_freq)) +
#   geom_point() +
#   geom_smooth(method='loess') +
#   geom_smooth(method='lm',color='red') +
#   theme_bw() +
#   ylab('variant frequency')
# ggMarginal(p, type = 'histogram')
```

In contrast, the correlation with viral load is much weaker. While it is strictly a positive correlation by assumption (Ne is assumed correlated with shedding duration, which is correlated with sampling time), selection rate is less sensitive to viral load than the time that selection has been allowed to operate before sampling.

```
p=ggplot(sim_data,aes(x=log10(Ne),y=var_freq)) +
  geom_point() +
  geom_smooth(method='loess') +
  # geom_smooth(method='lm',color='red') +
  theme_bw() +
  ylab('variant frequency')
ggMarginal(p, type = 'histogram')
```

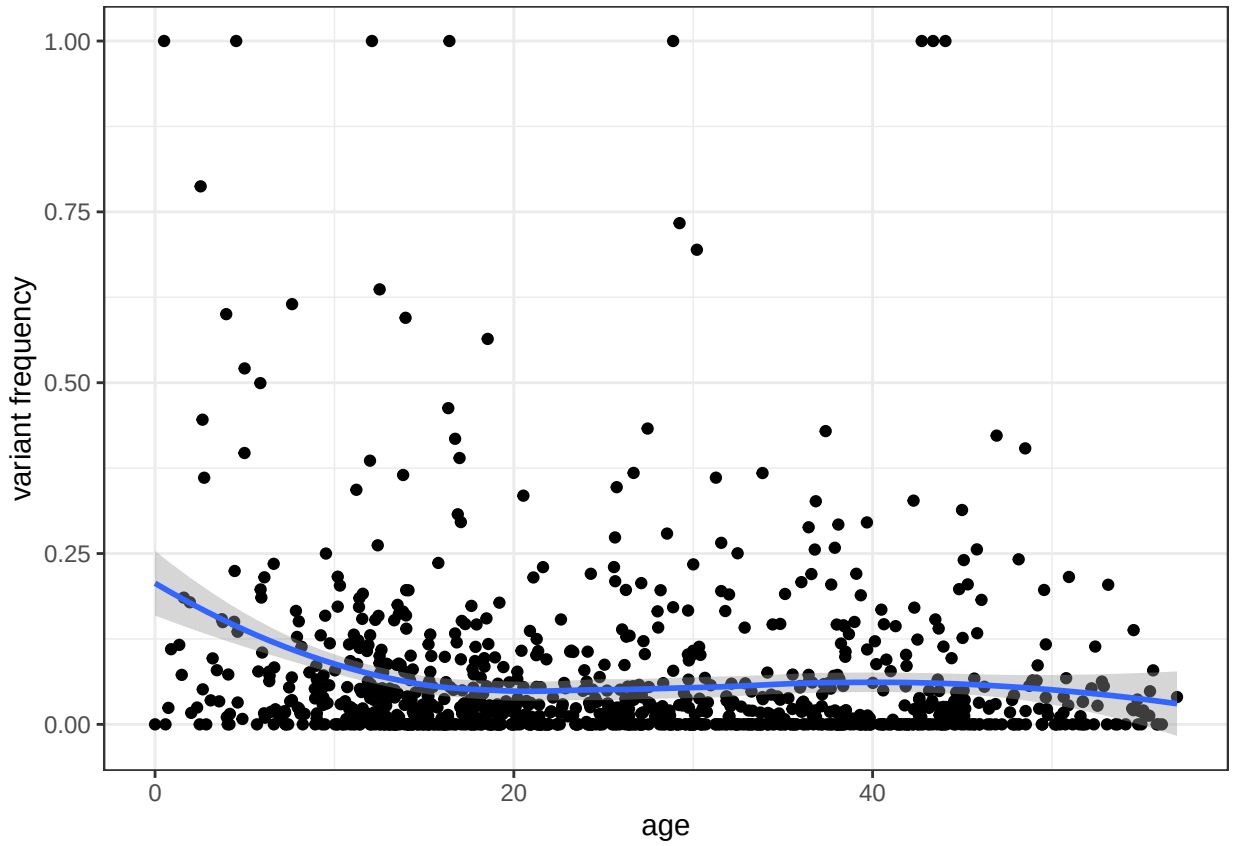


Taken together, this shows the basic hypothesis that longer infections will tend to be where you find variants.

And finally, let's look at how the association of shedding duration and age induces a superficially paradoxical pattern that selection appears to be stronger in younger children.

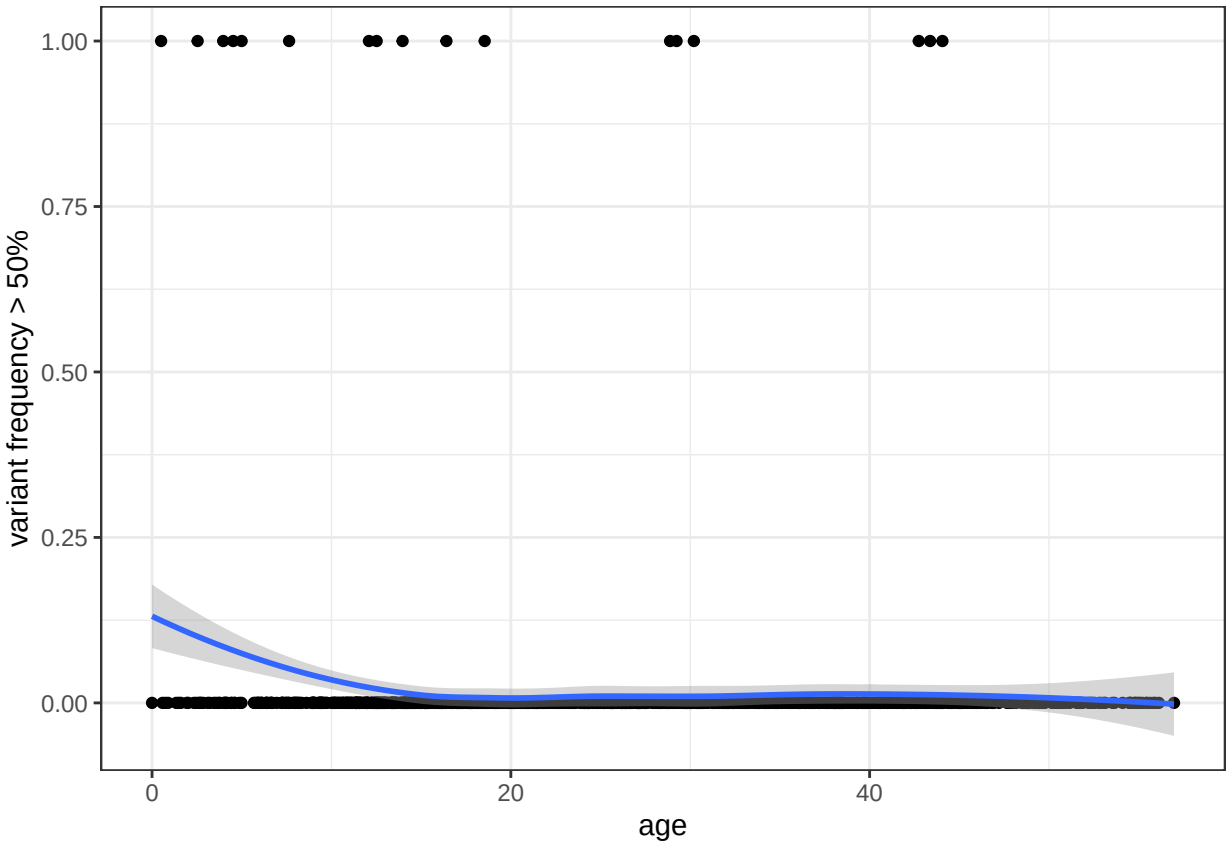
First, we look at variant fraction vs age.

```
ggplot(sim_data,aes(x=age,y=var_freq)) +
  geom_point() +
  geom_smooth(method='loess') +
  ylab('variant frequency') + theme_bw()
```



and here is consensus sequence variant vs age.

```
ggplot(sim_data,aes(x=age,y=as.numeric(var_freq>0.5))) +
  geom_point() +
  geom_smooth(method='loess') +
  ylab('variant frequency > 50%') + theme_bw()
```



This qualitatively reproduces the result I saw in the talk, without invoking any differential immune selection dynamics by age (which would manifest as associations between the selection coefficent s and age).

Summary

This quick modeling exercise demonstrates that known associations of age and shedding duration, and the survival bias the induces a correlation between younger ages and later sampling times during infection, can explain the superficially paradoxical observation of higher variant frequnencies at an antigenic site with younger ages, without invoking an age-dependent differential immune selection mechanism.

There of course also could be such a mechanism, but evidence for it requires finding a signal specific to immune selection that cannot be explained by shedding dynamics and sampling alone.

That said, this result is still interesting, as it indicates that for any sites in OC43 that are under less purifying selection during within-host replication than between-host transmission, variants are more likely to arise in children than adults.

This model could be fit to the real data to better represent the observations and allow a framework to explore any additional effects. In particular, the figure I recall seeing in the talk had a much stronger association with age. Which, if the model cannot fit it with realistic, data-informed parammters, that would indicate an additional need for the selection coefficient to be larger with younger ages.