# SARS-CoV-2 Herd Immunity Accounting for Population Structure

Santiago M. Castro Dau

Bayesian Phylodynamics, Spring 2021

## Introduction

### Overview

Herd immunity is achieved against a disease when a large enough portion of the susceptible population becomes immune to the disease, either by vaccination or contagion, thereby opposing the flow of the disease through the population. For SARS-CoV-2 the critical portion of the population ($p_{crit}$) was estimated to be around 60 to 70 percent but it is unclear if this estimate took population structure into account (Aschwanden 2021). In a structured population setting there would be a critical value $p_{crit,\,i}$ for each subpopulation describing the number of individuals that should be immunized in each subpopulation to bring the effective reproductive number ($R_e$) below 1. Naturally there could be multiple combinations of $p_{crit,\,i}$'s that could then accomplish this. Therefore the existence of subpopulations that are more prone to contract and/or spread SARS-CoV-2 would potentially change the way in which we design vaccination campaigns since the immunization of hyper-spreaders would have a larger effect in bringing down $R_e$ than hypo-spreaders. The goal of this project was to estimate the age-class specific reproductive numbers in Switzerland using BEAST2 and use them to estimate different combinations of $p_{crit,\,i}$ that would result in herd immunity (Bouckaert 2019).

### The Data

The sequence data was obtained from the GISAID data set, where we only considered samples from Switzerland from the 15th of November 2020 to the 17th of January 2021; a period in which restrictions were kept relatively constant (Oxford Stringency Index of 55 from the 6th of November to the 21st of December 2020 and 68 until the 17th of January) (Huisman et al. 2020). In order to filter out infection from before this restriction period we pushed our window of interest 9 days after the start of the restrictions on the 6th of November 2020. Also very importantly, this period is prior to the start of the vaccination campaign so we can observe the dynamics of the disease without the effect of this intervention.

The data comprised 201 sequences which we grouped into three age classes, young (30 and below), adult (between 30 and 59) and old (60 and above). The number of observations in each age class along with some summary statistics regarding the age distribution of each class are shown in table 1.

Table 1: Summary statistics regarding the age of the individuals and age groups considered.

| Country | Number of observations | Mean | Median | Standard Deviation | Min | Max |
|---|---|---|---|---|---|---|
| young | 52 | 20.17 | 21 | 6.76 | 6 | 29 |
| adult | 102 | 44.01 | 45 | 8.92 | 30 | 59 |
| old | 47 | 74.74 | 74 | 9.81 | 60 | 96 |

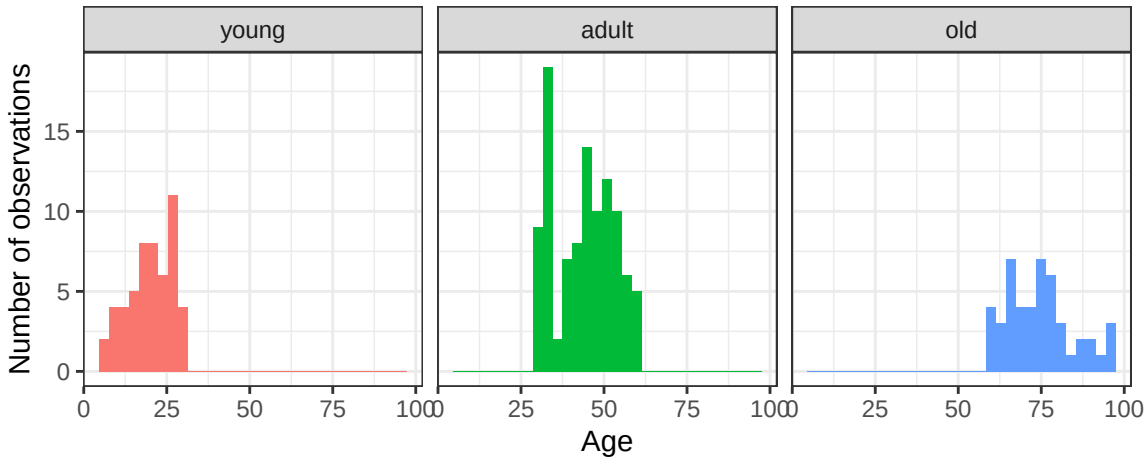


Figure 1: Frequency of observations according to age, subdivided by age group. Bin width of 3 year.

## Methods

All of the code mentioned below is available in this GitHub repository.

### Preprocessing

The first step involved acquiring the sequences of interest, annotating and aligning them. Because the age annotated sequences in GISAID are mixed with non annotated sequences for any given period we started by downloading all the available sequence for the period of interest, and filtering out observations that were not age annotated using a costume R script. Once we had all the ids of the age annotated sequences we downloaded them into the same `fasta` file. Then we changed the id of each sequence in the `fasta` file such that it included the GSAID id, it's date and the age group it belonged to. We proceeded to align the sequences using the `muscle` library (Edgar 2004). Both the annotation and the alignment were performed with the same costume R script.

### BEAST 2

To obtain the age-class specific $R_e$ estimates from our aligned data we proceeded to do bayesian phylodynamic inference BEAST 2 (Bouckaert 2019). First we loaded a multitype birth death template in BEAUti, loaded are sequences and then proceeded to specify the substitution model, clock model and priors. Everything that is not explicitly mentioned below was left in the default setting. Furthermore the `xml` file is available in the aforementioned GitHub repository.

For the site model we specified 4 gamma categories, set the proportion of invariant sites to 0.86 and enabled the estimate option for the later. For the substitution model we choose the Hasegawa-Kishino-Yano model as it accounts for most of the biases that can arise from analyzing nucleotide sequences (Hasegawa, Kishino, and Yano 1985; Barido-Sottani et al. 2017). For the clock model we choose a strict clock with a clock rate of 0.0098 as reported in the literature for SARS-CoV-2 (Dorp et al. 2020). For the priors we chose the following:

- $R_0$'s upper bound was set to 10.
- Log normal distribution for the becoming un-infectious rate with mean in real space equal to 36 (average 10 days).
- Log normal distribution for the clock rate with mean in real space equal to 0.0098 (Dorp et al. 2020).
- Log normal distribution for the sampling proportion with mean in real space equal to 0.0007.

The sampling proportion was derived the following way. According to the OWID data set there were `238,093` new confirmed SARS-CoV-2 cases reported in the period of interest (Hannah Ritchie and Roser 2020). Assuming that these new cases only account for 80% of the total cases (e.g. there is another 20% that does not get recorded, either because their are asymptomatic or for some other reason) we can estimate that

there were in total 285,712 new infection during this period. As we only have 201 sequences our sampling proportion should be 0.0007. The assumption that only 80% of the total cases are recorded is not supported by evidence, it is merely guided by the author's intuition.

The chain length was increased to $10^8$ and the frequency of the log's to $10^6$. Because 20 individual runs were performed the output file names were altered such that they would include the seed number. The analysis was run on ETH Zurich's Euler Cluster for 120 hours, however this amount of time was not enough to reach a chain length of $10^8$ so the analysis was restarted two more times using the `-resume` setting available on BEAST 2 (Bouckaert 2019). In summary the analysis was run three separate times, each terminating when $10^8$ iterations reached or when 120 hours of run time were reached, whichever happened first. As a result and because not all chains have the same speed the 20 individual runs have chain length between TODO and TODO.

**Exploring the critical proportions**

Once the age-class specific $R_e$ estimates were obtained, we proceeded to explore the combinations critical portions $p_{crit,\,i}$'s for which the $R_e$ would fall below 1. The effective reproductive number of a heterogeneous population ($R_e^{het}$) is given by a weighted sum of the class specific effective reproductive numbers (shown for three different risk classes A, B, C). The weights in this sum are determined by the fraction of infected individuals in each subpopulation.

$$R_e^{het} = \alpha_A R_e^A + \alpha_B R_e^B + \alpha_C R_e^C$$
$$\alpha_i = \frac{I_i}{I_A + I_B + I_C}$$

Ideally we would estimate these $\alpha_i$'s from all the newly reported cases in the period of interest but since such data is not readily available we will hope that our sequences represent a unbiased random sample from the underlying newly reported cases distribution and use the number of annotated sequences as a proxy to estimate these weights. Consequently $\alpha_{young} \approx 0.26$, $\alpha_{adult} \approx 0.51$ and $\alpha_{old} \approx 0.23$.

Interestingly we know that for this period the whole population $R_e^{het}$ was around 1 during this period, so we can check how acceptable this approximation is once we have our estimates for the age-class specific effective reproductive numbers (Huisman et al. 2020). Namely,

$$R_e^{het} = 0.26 R_e^{young} + 0.51 R_e^{adult} + 0.23 R_e^{old} \approx 1$$

On the other hand we know that the effective reproductive number of a vaccinated population ($\hat{R}_e$) is simply the original effective reproductive number times the fraction of susceptible left after vaccination.

$$\hat{R}_e = R_e(1 - p * E)$$
$$R_e^{het} = \alpha_A R_e^A(1 - p_A * E) + \alpha_B R_e^B(1 - p_B * E) + \alpha_C R_e^C(1 - p_C * E)$$

Here $p_i$ is the fraction of vaccinated people in a specific age-class relative to that age-class subpopulation times the efficacy of the vaccine $E$. We will assume $E$ to be equal to 0.95 (the reported efficacy of the Moderna and Pfizer-BioNTech vaccines) (Katella 2021). Finally, with this last expression we can play around with different values of $p_i$ to analyze combinations of proportions for each subpopulation which could bring the $R_e^T$ to a value below 1. Because we expect the $R_e^{het}$ to already be around 1 there probably won't be a lot of margin to move the critical proportions around before $R_e^{het}$ is already under 1.

## Results

### Convergence

### $R_e$ estimates

### Case 1

### Case 2

## Conclusion

## Sources


Aschwanden, Christie. 2021. "Five Reasons Why Covid Herd Immunity Is Probably Impossible." Nature. 2021. https://www.nature.com/articles/d41586-021-00728-2.

Barido-Sottani, Joëlle, Veronika Bošková, Louis Du Plessis, Denise Kühnert, Carsten Magnus, Venelin Mitov, Nicola F. Müller, et al. 2017. "Taming the BEAST—A Community Teaching Material Resource for BEAST 2." *Systematic Biology* 67 (1): 170–74. https://doi.org/10.1093/sysbio/syx060.

Bouckaert, Timothy G. AND Barido-Sottani, Remco AND Vaughan. 2019. "BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis." *PLOS Computational Biology* 15 (4): 1–28. https://doi.org/10.1371/journal.pcbi.1006650.

Dorp, Lucy van, Damien Richard, Cedric C. S. Tan, Liam P. Shaw, Mislav Acman, and François Balloux. 2020. "No Evidence for Increased Transmissibility from Recurrent Mutations in Sars-Cov-2." *Nature Communications* 11 (1): 5986. https://doi.org/10.1038/s41467-020-19818-2.

Edgar, Robert. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (February): 1792–7. https://doi.org/10.1093/nar/gkh340.

Hannah Ritchie, Diana Beltekian, Esteban Ortiz-Ospina, and Max Roser. 2020. "Coronavirus Pandemic (Covid-19)." *Our World in Data*. https://doi.org//url%7Bhttps://ourworldindata.org/coronavirus%7D.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano. 1985. "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial Dna." *Journal of Molecular Evolution* 22 (2): 160–74. https://doi.org/10.1007/BF02101694.

Huisman, Jana, Jérémie Scire, Daniel Angst, Richard Neher, Sebastian Bonhoeffer, and Tanja Stadler. 2020. "Estimation and Worldwide Monitoring of the Effective Reproductive Number of Sars-Cov-2," November. https://doi.org/10.1101/2020.11.26.20239368.

Katella, Kathy. 2021. "Comparing the Covid-19 Vaccines: How Are They Different?" Yale Medicine. 2021. https://www.yalemedicine.org/news/covid-19-vaccine-comparison.