A Supplementary Material

A.1 Sequencing Protocols for Viollier samples and HUG samples

We perform whole-genome sequencing of the Viollier samples in three facilities. The HUG samples are processed in one of them, the Health 2030 Genome Center. The sequencing protocol for the samples sequenced at the Genomics Facility Basel and the Functional Genomics Center Zurich is described in [1]. Samples processed at the Health 2030 Genome Center used the Illumina COVIDSeq library preparation reagents following the protocol provided by the supplier [2]. These reagents are based on the ARTIC v3 multiplex PCR amplicon protocol [3]. When sufficient volume was available, 8.5ul of RNA extracted from patient nasopharyngeal samples were used in the cDNA synthesis step; if 8.5ul were not available, the maximum volume possible was used. Pooled libraries were sequenced on the Illumina NovaSeq 6000 using a 50-nucleotide pair-end run configuration. Post-sequencing library read de-multiplexing was done using an in-house developed processing pipeline [4]. The downstream bioinformatics procedure to obtain consensus sequences is described in [1, 5].

A.2 Screening procedure at Dr Risch

Dr Risch medical laboratories used the Taqpath assay from Thermofisher for their diagnostic and recorded the S gene target failure (SGTF). SGTF samples are potential B.1.1.7 variants, as the B.1.1.7 variant causes a SGTF due to its deletion at positions 69-70 in the spike protein. Further, Dr Risch medical laboratories screen their samples for the 501Y mutation by a variant-specific PCR test. If a sample is identified as a potential VoC by these procedures, it was initially sent for whole-genome sequencing to the University Hospital Basel in order to confirm the B.1.1.7 variant. The sequencing protocol is described in [6]. For recent samples, the confirmation may still be outstanding or not conducted due to B.1.1.7 now being dominant. However, since typically a SGTF plus a 501Y mutation corresponds indeed to a B.1.1.7 variant, we consider these samples as B.1.1.7 variants even when whole genome sequencing confirmation is lacking.

A.3 Estimating the transmission fitness advantage of a new variant

In what follows, we define two models describing the dynamics with which a new variant with a transmission fitness advantage spreads in a population. The first model is based on the assumption of discrete-time, while the second model is based on the assumption of continuous-time. Both models have been considered extensively in the literature to estimate fitness advantage (see e.g. [7, 8]). While in an epidemic, a generation does not end after a fixed time span (discrete-time model), generations are typically less variable than modeled under an exponential distribution (continuous-time model). Thus we view the two models as two extremes. We provide estimates based on both models and suggest that the true parameter may be anywhere within the ranges spanned by the two models. In the next sections, we provide details of how we estimate the transmission fitness advantage of B.1.1.7 based on daily data of the total number of samples and B.1.1.7 samples under these two models.

A.3.1 Discrete time model

We call X the common (non-B.1.1.7) variants and Y the B.1.1.7 variant. The process starts in generation 0 with x_0 cases caused by variant X and y_0 cases caused by variant Y. Let the number of cases in generation n be $x(n)$ and $y(n)$ for variants X and Y.

Let the reproductive number R_d of variant X in generation n be $R_d(n)$. Let the transmission advantage of variant Y be f_d. Then the reproductive number of Y in generation n is $(1+f_d)R_d(n)$. Thus, we assume a multiplicative fitness advantage.

We have $x(n) = x(0) \times R_d(0)R_d(1)\ldots R_d(n-1)$ and $y(n) = y(0) \times R_d(0)R_d(1)\ldots R_d(n-1)(1+f_d)^n$. If R_d is constant through time, we have

$$x(n) = x(0)R_d^n \text{ and } y(n) = y(0)((1+f_d)R_d)^n.$$

Let the proportion of variant Y at generation n be $p(n)$. We have,

$$\begin{aligned} p(n) &= \frac{y(n)}{x(n)+y(n)} = \frac{y(0)(1+f)^n}{x(0)+y(0)(1+f)^n} \\ &= \frac{p(0)(1+f_d)^n}{1-p(0)+p(0)(1+f_d)^n} \\ &= \frac{1}{1+(1+f_d)^{-n}\left(p(0)^{-1}-1\right)}. \end{aligned}$$

Thus, $p(n)$ is the logistic function. It does not depend on R_d.

If we write time in days t rather than generations n and assume a generation time of g days, we get

$$p(t) = \frac{1}{1+(1+f_d)^{-t/g}\left(p(0)^{-1}-1\right)}. \tag{1}$$

We now switch our parameterization to the more common

$$p(t) = \frac{1}{1 + e^{-a(t-t_0)}} \tag{2}$$

for parameter estimation from daily data. Parameter a is the logistic growth rate and parameter t_0 the sigmoid's midpoint.

The two free parameters, a and t_0, are related to the two free parameters in Equation 1, f_d and $p(0)$, as follows:

$$a = \frac{\ln(1 + f_d)}{g}, \quad t_0 = g\ \frac{\ln(p(0)^{-1} - 1)}{\ln(1 + f_d)}.$$

In particular, we get $f_d = e^{ag} - 1$.

A.3.2 Continuous time model

In continuous-time, instead of R_d and generation time g, we define the transmission rate β and the recovery rate μ. Under this model, the reproductive number is $R_c = \beta/\mu$. Further, since an individual in the discrete model recovers after a generation of duration g (during which they left R_d offspring), we note that g is related to the expected time to recovery $1/\mu$ in the continuous model, and in fact assume $g = 1/\mu$ in what follows. Again our initial numbers of the variants X and Y are $x(0)$ and $y(0)$. Calendar time is denoted by a continuous parameter t. We then have in expectation,

$$x(t) = x(0)e^{(\beta-\mu)t}. \tag{3}$$

We note that $\beta - \mu$ is coined the Malthusian growth parameter [8].

Further, we again assume that variant Y has a transmission fitness advantage of f_c, with transmission rate $\beta(1 + f_c)$ and recovery rate μ. The population size of the variant at time t is thus

$$y(t) = y(0)e^{(\beta(1+f_c)-\mu)t}. \tag{4}$$

The proportion of the variant in the population at time t is

$$\begin{aligned} p(t) &= \frac{y(0)e^{(\beta(1+f_c)-\mu)t}}{x(0)e^{(\beta-\mu)t} + y(0)e^{(\beta(1+f_c)-\mu)t}} \\ &= \frac{p(0)e^{\beta f_c t}}{1 - p(0) + p(0)e^{\beta f_c t}} \\ &= \frac{1}{1 + e^{-\beta f_c t}(p(0)^{-1} - 1)} \end{aligned}$$

where we again recognize the logistic function. We turn again to the more common parameterization,

$$p(t) = \frac{1}{1 + e^{-a(t-t_0)}} \tag{5}$$

where we thus have $f_c = \frac{a}{\beta}$.

The reproductive number is $R_c = \beta/\mu$ and the mean time to recovery, $1/\mu$, is equaled to g. Then, $\beta = R_c/g$. Thus,

$$a = \frac{f_c R_c}{g}, \quad t_0 = g\frac{\ln(p(0)^{-1} - 1)}{f_c R_c}.$$

In particular, we have $f_c = \frac{ag}{R_c}$. Note that the estimated fitness advantage under this model depends on the reproductive number R_c and is thus changing if R_c is changing through time.

A.3.3 Connection between discrete and continuous time

The discrete and continuous models are very similar. Both have the intitial conditions $x(0)$ and $y(0)$. For the dynamics, the discrete model has parameters R_d and g while the continuous model has parameters β and μ. We have $R_c = \beta/\mu$ and we further assumed that $g = 1/\mu$ ($1/\mu$ is the expected time until recovery in the continuous setting while g is the time to recovery in the discrete setting). The different parameterizations of fitness advantage are coined f_c and f_d. We now determine how f_c and f_d are related.

To compare the two models, we now assume that their overall dynamics for the variant X are the same. After n generations of duration g, we have for variant X,

$$\begin{aligned} & x(0)e^{(\beta-\mu)ng} = x(0)R_d^n \\ \iff & e^{(\beta-\mu)g} = R_d \\ \iff & R_d = e^{R_c - 1} \end{aligned}$$

Using a Taylor expansion for β/μ close to 1, we obtain that indeed $R_d = R_c$.

For the two models to produce the same growth also for variant Y, we require,

$$\begin{aligned}
& y_0 e^{(\beta(1+f_c)-\mu)ng} = y_0((1+f_d)R_d)^n \\
\Longleftrightarrow & (\beta(1+f_c)-\mu)g = \ln((1+f_d)R_d) \\
\Longleftrightarrow & (\beta-\mu)g + f_c\beta g = \ln(R_d) + \ln(1+f_d) \\
\Longleftrightarrow & f_c\beta g = \ln(1+f_d)) \\
\Longleftrightarrow & f_c = \frac{\ln(1+f_d)}{\beta g} \\
\Longleftrightarrow & f_c = \frac{\ln(1+f_d)}{R_c}.
\end{aligned}$$

In the last step, we make use of $R_c = \beta/\mu$ and $g = 1/\mu$.

This is equivalent to $f_d = e^{R_c f_c} - 1$. Using a Taylor expansion we get $f_d = 1 + R_c f_c + O((R_c f_c)^2) - 1$ and thus $f_d = R_c f_c$ for small $R_c f_c$.

A.3.4 Maximum likelihood parameter estimation

Next we explain how we estimate a and t_0 of the logistic functions (Eqn. 2 and 5) from our data using maximum likelihood. We consider that we have data at times $t_1, ..., t_d$. At time t_i, we obtained n_i samples, where n_i is fixed, non-random.

We assume that the true number of B.1.1.7 variants at time t_i is a random variable, K_i which is binomially distributed with parameter $p(t_i)$, i.e.

$$K_i \sim \mathcal{B}\left(n_i, p(t_i)\right), \qquad \text{where } p(t_i) = \frac{e^{a(t_i-t_0)}}{1+e^{a(t_i-t_0)}} \quad .$$

In particular, we assume here a deterministic logistic growth model for the increase in the proportion of variant Y (Eqn. 2 and 5), on top of which only the drawing process is random. This model simplifies naturally to a very popular logistic regression. This is an instance of a Generalized Linear Model, where the natural parameter of the binomial distribution is a linear function of predictors, the only predictor considered here being the time t.

We use the *Python* library *statsmodel* [9] to recover maximum likelihood estimates (MLEs) and confidence intervals. The confidence intervals are based on an asymptotic Gaussian distribution for the parameters of the logistic regression fitted to our data, i.e. the fixed values $t_1, \ldots, t_d$, $n_1, \ldots, n_d$ as well as the numbers of samples at each time point being the variant B.1.1.7, $k_1, \ldots, k_d$. Given the large number of sequences – we have at least 100 samples per region – the use of an asymptotic Gaussian approximation is justified. Parameters a, t_0, f_d, f_c as well as the proportions of variant B.1.1.7 $p(t)$ through time are simple transformations of the parameters of the logistic regression. Their MLEs are the same transformations applied to the MLEs of the logistic regression parameters. The difference between the MLEs and the true parameters are again Gaussian, with a covariance matrix found by applying the delta method. This is used to construct confidence intervals for all these quantities. We used the default fitting procedure provided in the *statsmodel* package. This procedure reported convergence for all analyses.

A.4 Estimation of the effective reproductive number

We use the number of confirmed cases per day from the Federal Office of Public Health, Switzerland, for 14 December 2020 to 11 March 2021. Then, for each day, we estimate the number of B.1.1.7 variants by multiplying the total number of confirmed cases by the proportion of B.1.1.7 in our dataset (Viollier or Risch). We then estimate an effective reproductive number of the B.1.1.7 variant and of the non-B.1.1.7 variants using these data. For this estimation, we use the method developed in [10]. This method consists of two main parts: first, the observed case data is related to the corresponding time series of infections. We smooth the observations using LOESS smoothing to remove weekend effects. Then, we deconvolve with the delay of infection to symptom onset (gamma distributed with mean 5.3 and sd 3.2) and the delay from symptom onset to case confirmation (gamma distributed with mean 5.5 and sd 3.8). Second, we estimate the effective reproductive number from the time series of infection incidence using EpiEstim [11]. The reported point estimate is the estimate on the original case data. To account for uncertainty in the observation process, the observed daily case incidences are additionally bootstrapped 1000 times, resulting in an ensemble of alternative case incidence time series and corresponding estimated effective reproductive numbers. These are used to construct the 95% confidence interval around the effective reproductive number, and to calculate the standard deviation of the ratios of effective reproductive number estimates (see below).

We perform the estimation of the reproductive number in two different ways. First, we estimate smooth changes in the reproductive number, by estimating it across the entire time series using a 3-day sliding window. Second, we assume the reproductive number was constant during time intervals in which the non-pharmaceutical interventions did not change. Since 18 January 2021, Switzerland has implemented a set of tighter measures (in particular, shops are closed and the size of gatherings is restricted to five people [12]). Thus we fix the reproductive number to be constant between 01 January and 17 January 2021. Then the reproductive number is allowed to change and again fixed to be constant from 18 January 2021 onwards.

To compare the effective reproductive number R of the B.1.1.7 variant (Y) to that of non-B.1.1.7 variants (X), we take the ratio $\rho = \frac{R_Y}{R_X}$ at every time point. The standard deviation of this ratio σ_ρ was found through Gaussian error propagation of the standard deviation of the individual R estimates (σ_X, σ_Y):

$$\sigma_\rho = \sqrt{\frac{1}{(R_X)^2}\sigma_Y^2 + \frac{(R_Y)^2}{(R_X)^4}\sigma_X^2}.$$

A.5 Discrepancy for Ticino and Lake Geneva

The discrepancy for Ticino and Lake Geneva is not surprising: they had a reproductive number which was different from the national reproductive number in the first half of January. For Ticino, the empirical case numbers drop faster than the model, which is in line with a lower reported reproductive number compared to the national level[13]. For the Lake Geneva region, the empirical case numbers drop slower than the model, which is in line with a higher reported reproductive number compared to the national level[13]. For all regions but Ticino, we have enough data to estimate a reproductive number for the non-B.1.1.7 variants for 01 January-17 January 2021. While for Switzerland, we obtained a point estimate of 0.83, the point estimates for all regions but Lake Geneva are between 0.81-0.83. Thus using the point estimate for all of Switzerland for the regional plots - with the exception of Lake Geneva and Ticino - in Fig. 5 is justified. For Geneva, we obtain a point estimate of 0.88. We use this point estimate in a Lake Geneva specific model (Fig. S1). Again we observe that the total number of confirmed cases dropped recently faster compared to the model. For a discussion on the discrepancy see main text.

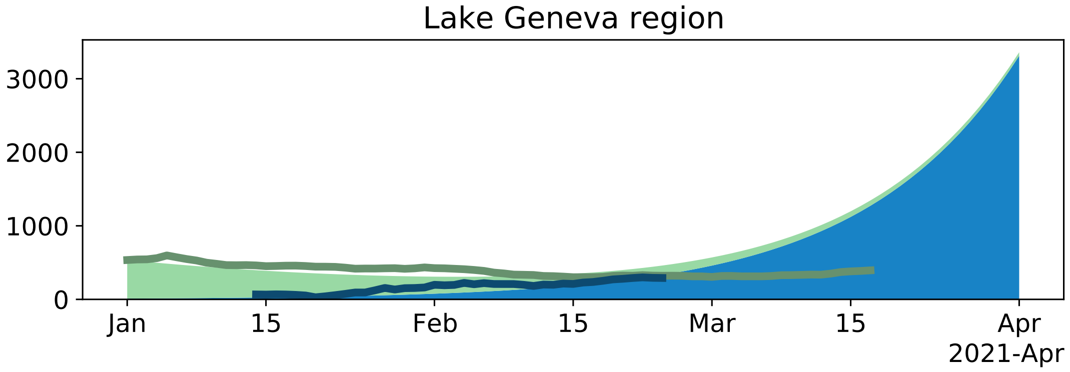

Figure S1: Change in the number of B.1.1.7 variants and in the number of all cases through time for the Lake Geneva region. For details see legend of Fig. 4. Compared to Fig. 4, we here use the average reproductive number estimated for non-B.1.1.7 in Geneva for the time period 01 January 2021-17 January 2021. The transmission fitness advantage is calculated based on this reproductive number and the estimate of the growth rate a for the Lake Geneva region.

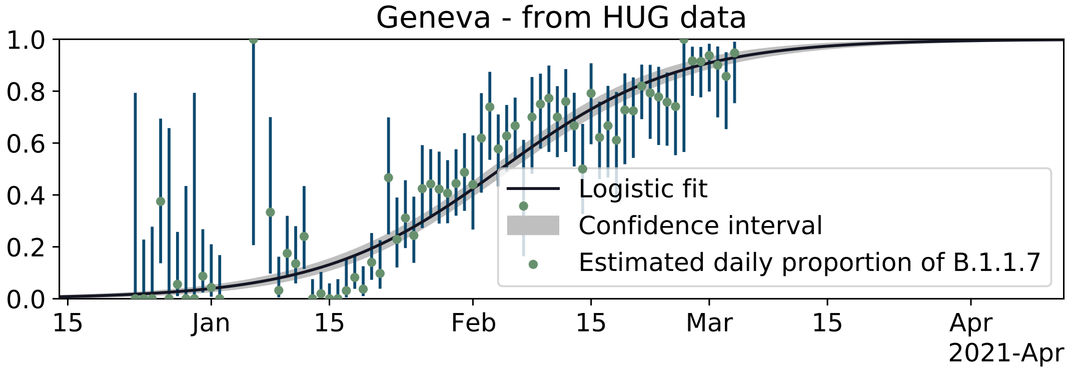

Figure S2: Logistic growth of frequency of B.1.1.7 in the Lake Geneva region based on HUG data.

A.6 Additional Tables

Grossregion	Total confirmed cases	Sequenced	Proportion
Central Switzerland	17663	503	2.85%
Espace Mittelland	35933	2948	8.20%
Nordwestschweiz	24275	2716	11.19%
Tessin	9442	119	1.26%
Zurich	33615	1820	5.41%
Lake Geneva	36515	807	2.21%
Ostschweiz	26722	859	3.21%
Total	184165	9772	5.31%

Table S1: The proportion of sequenced cases out of all cases for the Viollier dataset.

Grossregion	Total confirmed cases	Sequenced	Proportion
Central Switzerland	17663	321	1.82%
Espace Mittelland	35933	3893	10.83%
Nordwestschweiz	24275	1135	4.68%
Tessin	9442	113	1.20%
Zurich	33615	1090	3.24%
Lake Geneva	36515	405	1.11%
Ostschweiz	26722	4594	17.19%
Total	184165	11551	6.27%

Table S2: The proportion of characterized cases out of all cases for the Risch dataset.

A.7 GISAID Accession Numbers

The used sequences were, if fulfilling the quality criteria, uploaded to GISAID. The GISAID accession numbers are in the files supplementary_material_a7_d-bsse_gisaid_ids.txt and supplementary_material_a7_hug_gisaid_ids.txt which are attached to this paper.

References

[1] Nadeau SA, Vaughan TG, Sciré J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. Proceedings of the National Academy of Sciences. 2021;118(9).

[2] Illumina COVIDSeq Test;. Available from: https://emea.illumina.com/products/by-type/ivd-products/covidseq.html.

[3] ARTIC v3 multiplex PCR amplicon protocol;. Available from: https://artic.network/.

[4] Health 2030 Genome Center github;. Available from: https://github.com/health2030genomecenter.

[5] Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, Beerenwinkel N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. Bioinformatics. 2021 01. Btab015. Available from: https://doi.org/10.1093/bioinformatics/btab015.

[6] Stange M, Mari A, Roloff T, Seth-Smith HM, Schweitzer M, Brunner M, et al. SARS-CoV-2 outbreak in a tri-national urban area is dominated by a B.1 lineage variant linked to a mass gathering event. PLOS Pathogens. 2021 03;17(3):1-20. Available from: https://doi.org/10.1371/journal.ppat.1009374.

[7] Chevin LM. On measuring selection in experimental evolution. Biology letters. 2011;7(2):210-3.

[8] Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. Science. 2021;372(6538).

[9] Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference; 2010. .

[10] Huisman JS, Scire J, Angst DC, Neher RA, Bonhoeffer S, Stadler T. Estimation and Worldwide Monitoring of the Effective Reproductive Number of SARS-CoV-2. medRxiv. 2020 Nov:2020.11.26.20239368.

[11] Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology. 2013 Nov;178(9):1505-12.

[12] Coronavirus: Federal Council extends and tightens measures;. Available from: https://www.admin.ch/gov/en/start/documentation/media-releases/media-releases-federal-council.msg-id-81967.html.

[13] Real-time estimates of the reproductive number for SARS-CoV-2;. Available from: https://ibz-shiny.ethz.ch/covid-19-re-international/.