Inferring SARS-CoV-2 variant within-host kinetics

Baptiste Elie[1,*], Emmanuel Lecorche[2], Mircea T. Sofonea[1],
Sabine Trombert-Paolantoni[2], Vincent Foulongne[3], Jérémie Guedj[4]
Stpéhanie Haim-Boukobza[2], Bénédicte Roquebert[2], Samuel Alizon[1,*]

[1] MIVEGEC, CNRS, IRD, Université de Montpellier, Montpellier, France
[2] Cerba Laboratory, Saint Ouen L'Aumône, France
[3] Laboratoire de Virologie, CHU de Montpellier, France
[4] Université de Paris, INSERM, IAME, F-75018 Paris, France
[*] Corresponding authors: baptiste.elie@ird.fr samuel.alizon@cnrs.fr

Abstract

SARS-CoV-2 variants are causing epidemic rebounds in many countries. By analyzing longitudinal cycle threshold (Ct) values from screening tests in the general population and hospitals, we find that infections caused by variant lineages have higher peak viral load than wild type lineages and, for the B.1.1.7 lineage, have a longer infectious period duration. Linking within-host kinetics to transmission data suggests that infections caused by variants have higher transmission potentials and that their epidemiological fitness may depend on the demography of the host population.

keywords COVID-19 | statistical modelling | virus load | variants of concern

Introduction

The end of the year 2020 has seen the identification of three SARS-CoV-2 lineages causing phenotypically different infections from 'wild-type' lineages. These are referred to as 'Variants of Concern' (VOC) by the World Health Organisation (WHO). The first VOC (V1) belongs to the Pango lineage B.1.1.7 and has been shown to be more contagious [1, 2] and more virulent [1, 3]. The other two VOCs (V2 and V3) belong to lineages B.1.351 and P.1. They also seem to be more contagious than wild type lineages [4] but also with a potential to evade host immunity from previous infection [5, 6]. These three VOCs are associated with deadly epidemic rebounds in several countries, which has led to their close monitoring through full genome sequencing but also targeted Reverse-Transcription Polymerase Chain Reaction (RT-PCR) screening. The latter is less precise than the former but more affordable, allowing for wider testing [7].

RT-PCR have been reported to exhibit lower cycle threshold (Ct) values for V1 than for wild type infections [1, 8]. This suggests that VOC may be causing infections with higher virus loads but these analyses only included a single time point per individual. Here, we analyze longitudinal follow-up of 8,763 individuals to investigate potential changes in the virus kinetics in infections caused by variants.

The field of within-host kinetics has grown focusing mainly on chronic infections, but some studies consider acute infections [9]. In the case of SARS-CoV-2, numerous studies used Ct values as a proxy for virus load to report temporal variations within individuals. Some of these performed statistical analyses on longitudinal data from hospitalized patients [10], health workers [11], and experimental infections in non-human primates [12]. These studies show that viral kinetics are associated with the infection and the detection probability. In addition to the number of individuals followed, our current study stands out in two ways. First, we analyze data from the general population. Second, we compare infection kinetics between variant and wild-type strains.

Results

After screening the database, we identified 17,113 suitable samples from 8,006 individuals. The median age was 42 with an interquartile range (IQR) of [26,59]. A minority of samples originated from hospitals (6%). Most individuals (88%) were sampled twice and the median follow-up

duration was 8 days (IQR [6-12] days). Most samples originated from the Ile-de-France region (64%) and were caused by the V1 variant (73%), which reflects the state of the French epidemic at the time where the samples were collected [13]. 21% of the tests were assigned to wild type (WT) strains and 6% to V2 or V3.

Virus load kinetics

We analysed Ct values using linear mixed models with a random effect for each patient and a variety of co-factors and interactions. The model selected using the AIC included the following co-factors: the age, the lineage, the interaction between the two, the hospitalization status, the infection day, and the interaction between the day and either the lineage, the hospitalization status and the age (Table 1).

Table 1: **Linear mixed model parameters affecting Ct values.** Only significant effects are shown in the table. The standard deviation of the random effect on the intercept is 2.39 [2.27, 2.49] and that of the residues is 4.24 [4.17, 4.30]. The notation a:b indicates an interaction between factors a and b. B-F-C stands for 'Bourgogne-Franche-Comté'.

Predictor		**Estimate**	**95% CI**
Intercept		23.4	(22.9,23.9)
day		1.19	(1.15,1.23)
strain	*V2/V3*	-1.27	(-2.21,-0.332)
hospital	*yes*	-1.01	(-1.46,-0.551)
region	*B-F-C*	-2.49	(-3.68,-1.3)
	Normandie	-1.01	(-1.3,-0.713)
	Nouvelle-Aquitaine	-1.3	(-1.9,-0.703)
	Occitanie	-2.63	(-3.17,-2.09)
date		0.0066	(0.000851,0.0124)
age:strain	*V1*	-0.014	(-0.0234,-0.00456)
day:strain	*V1*	-0.0481	(-0.0795,-0.0166)
day:hospital		-0.0725	(-0.125,-0.0201)
day:age		-0.00358	(-0.00415,-0.003)

The linear mixed model revealed significant differences in viral load dynamics between strains (Figure 1). Compared to wild type infections, the peak virus load appeared to increase with age for infections caused by V1 (Table 1). When using the French demographic age structure, we find a 1.04 Ct difference between V1 and wild-type infections, suggesting a higher virus load

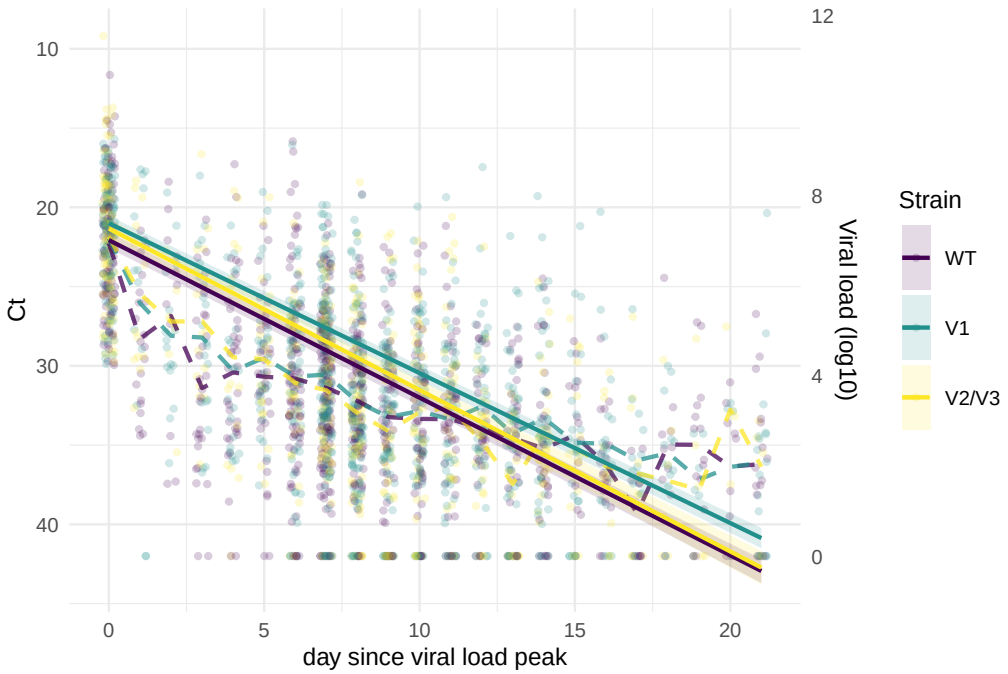

Figure 1: **Within-host SARS-CoV-2 Ct kinetics for three virus strains.** The dots represent the observed values and the dashed lines the daily mean value for each strain. The lines represent the linear model for an average patient (median age and not in a hospital setting). For each individual, day 0 is the day with the lowest Ct value.

in the former ($p<10^{-4}$, using the Tukey method and the emmeans R package). Ct values were lower in V2/V3 compared to the wild type with a Ct difference of 0.754 (p=0.0072). Furthermore, in infections caused by V1, the virus load decline rate was lower than in wild type infections. For V2/V3, this effect was not significant.

A potential bias is that large Ct values (> 37) could indicate the absence of the virus rather than a low viral load, therefore leading to a censoring effect. This can be seen in Figure 1, where the linear regressions (plain lines), overestimate the observed Ct (dashed lines). To address this issue, we performed survival analyses with Cox regression models, using the crossing of a value of 30 by the Ct as a termination event. This threshold is commonly used in diagnostics and is consistent with our mapping between Ct decline and infectiousness profiles (Figure 2A). We found a significant effect for variant V1 compared to the wild type strain (hazard ratio, HR, of 0.81, 95%CI: [0.75-0.88]), with a median increase in the duration of the decline phase of the virus load of 0.86 days. The variants V2/V3 did not show a significant difference from the wildtype (95% CI of the HR: [0.79-1.06]).

Our analysis also detected significantly higher peak viral loads and slower virus load decline rates in hospitalized individuals. This was confirmed by the survival analysis, which yielded

an increase of 1.46 days associated with samples from hospitals (HR = 0.71, 95%CI: [0.61,0.83]). Finally, the region of sampling and the sampling date were significantly associated with the Ct values. These are consistent with the fact that Ct values are known to vary depending on the facility that performed the sampling [14]. Furthermore, the association between the Ct and the sampling date can be explained via variations in epidemic temporal reproduction number [15].

These results are robust to the assumptions made regarding the dataset formatting, especially the inclusion or not of individuals above 80 years old, of hospitalized patients, or of patients with a long clinical follow-up, i.e. more than 21 days (results not shown).

Transmission potentials

To quantify the implications of these differences in kinetics at the population level, we performed a mapping between the decline of the daily infectivity after its peak (using the estimates from [16]) and the daily Ct value for the wild type strain inferred from the linear model. We found a significant correlation (R^2=95%, Figure 2A), which supports the hypothesis of a linear relationship between Ct and infectiousness and is consistent with both metrics being correlated to the log of the virus load. Using this mapping between Ct values and infectivity, we found that variant strains had higher transmission potentials compared to wild type strains (Figure 2B). For V1, this transmission advantage was more pronounced in countries with older populations such as France or Japan, whereas for V2/V3 it was slightly higher in countries with younger populations, such as Niger.

Discussion

Understanding the within-host kinetics of SARS-CoV-2 infections already already yielded original insights on infection virulence [10], or efficiency of screening strategies [11]. Here, we analyzed a large national dataset of longitudinal RT-PCR Ct values to test the hypothesis that epidemic rebounds associated with SARS-CoV-2 variants could be explained by phenotypic differences in the infections they cause.

A linear mixed model indicated that infections caused by SARS-CoV-2 variants have higher virus loads, with a significant age-dependence for strain V1. Furthermore, the temporal decrease in virus load was found to be slower when infections were caused by V1 instead of wild type

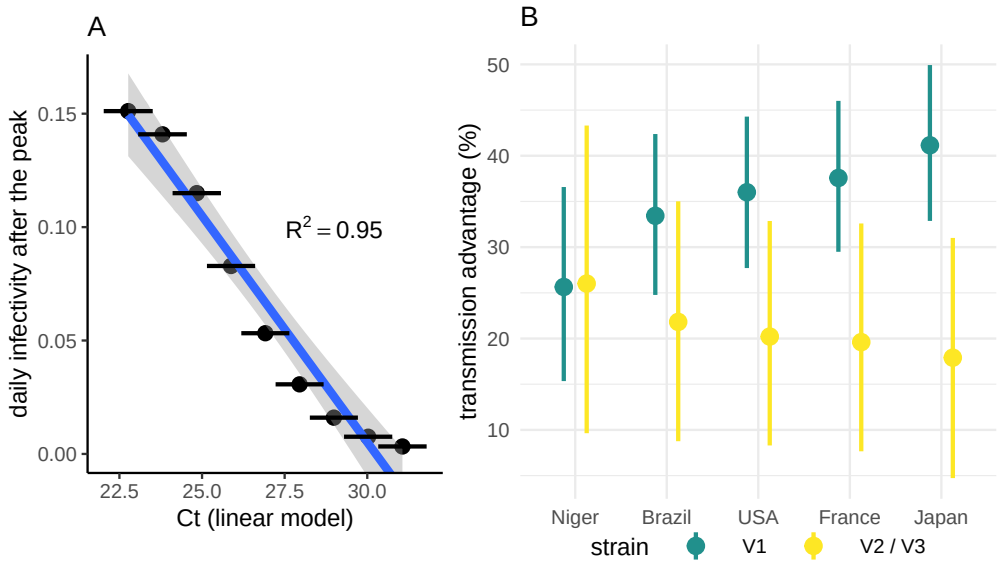

Figure 2: **Impact of within-host kinetics on infectiousness and transmission potential.** A) Correlation between infectiousness after the viral load peak and Ct values for the wild-type strain. B) Variant transmission advantage over the wild type for different countries. In B, we use the linear model parameter estimates and the output of the regression in A to compute transmission potentials. The bars indicate the 95% bootstrap quantiles.

strains. This result was confirmed using a survival analysis, which indicated a significant increase of 0.86 day. The results are consistent with unpublished results from a different cohort, which had fewer individuals with 2 or more samples (2,633 versus 8,006 here) but had self-reported symptom onset dates [17].

To further investigate the consequences of these variations in within-host kinetics at the epidemiological level, we first performed a correlation between the SARS-CoV-2 infectiousness profile and the estimated kinetics. This revealed a linear correlation between Ct value and daily infectiousness. Given that Ct values reflect logarithmic differences in virus load, this trend is consistent with earlier studies [18]. We then translated the estimated Ct kinetics into transmission potential profiles. This revealed that variants have a higher transmission potential than wild type strains. Furthermore, this advantage was found to be more pronounced for V1 in older populations (e.g. with a median advantage of 41% in Japan). For V2/V3, we found the opposite trend (e.g. with a median advantage of 26% in Niger).

Kinetics of samples collected in hospitals exhibited higher peak viral loads (i.e. lower Ct values), and we also found a longer period under a Ct of 30 in hospital settings (1.46 days), which is consistent with earlier studies [10]. Note that the other results were unaffected when we removed the data from the hospitals from the analysis.

A limitation of this analysis is that we do not have any indication regarding the date of the infection or of the symptom onset. This uncertainty prevents us from analyzing more mechanistic models with nonlinear mixed-effect models [10]. However, since we the nature of the virus causing the infection is unlikely to affect the number of days between infection and screening, we do not expect our assumption that the lowest Ct value corresponds to the peak viral load to introduce biases. Furthermore, the variant assessment for V2/V3 should be treated with caution because it only relies on the N501Y mutation. However, the assessment of the V1 variant is robust. Combining longitudinal Ct data and sequencing data [19] could be a way to improve the resolution of the analyses.

Another future extension of this approach will be to combine Ct data with serological data analyse the immune evasion properties of V2 and V3 variants [5, 6].

Material and methods

Data

This study was approved by the Institutional Review Board of the CHU of Montpellier and is registered at ClinicalTrials.gov under the identifier NCT04653844. The data used originates from more than 208,128 variant-specific RT-PCR tests performed in France between February 06 and April 14, 2021 on SARS-CoV-2 positive samples [8]. The assay used was IDTM 38 SARS-CoV-2/UK/SA Variant Triplex (ID SOLUTION) with probes targeting the spike Δ69-70 deletion and N501Y mutation, as well as a region in the N virus gene for control purposes. We use the Ct of the later in our analyses. Samples with both the deletion and the N501 mutation were considered to belong to the B.1.1.7 lineage (i.e. the V1 variant), samples with only the N501Y mutation to the B.1.351 or the P.1 lineage (i.e. variants V2 or V3), and sample without any of the two to wild type lineages. Although the assay cannot discriminate between V2 and V3, sequencing from the national health agency (Santé Publique France) indicate that V3 is marginal.

We identified individuals with multiple samples and made the following assumptions before performing our statistical analyses: i) for each individual, the day with the lowest Ct value was defined as day 0, ii) only samples collected up to day 21 were analysed, iii) to avoid bias in the variant assessment, samples with a Ct > 30 were treated as "uninterpretable", and all "uninterpretable" results were ignored in assessing the strain causing an infection, iv) individuals

were assumed to be infected with only one strain at a time and follow-ups with multiple strains being detected were discarded, and v) follow-ups consisting only of "uninterpretable" tests were removed. The exact steps of the data selection are shown in Supplementary Materials.

Statistical analyses

We analyzed the longitudinal data of Ct values with linear mixed models and used the R package lme4 to fit the restricted maximum likelihood parameters to the data. The response variable was the Ct value and the main effects included in our model comparison were the virus strain (V1, V2/V3, or wild type), the day (day 0 being that with the lowest Ct value), the hospitalization status, the age, the sampling region, and the sampling date. We also included interactions, based on the Akaike Information Criterion (AIC). For this, the AIC was computed on the corresponding linear mixed models, refitted to their maximal likelihood value. Differences in Ct values between populations from the linear model outputs were computed using the Tukey method and the emmeans R package.

To account for potential statistical biases associated with data right-censoring, we performed a survival analysis using a semi-parametric Cox model and the survival R package. We considered data with a Ct above 37 as right-censored.

Using Ct measures as a proxy for virus load has several limitations, especially in the case of Coronaviruses [20]. Here, we do not attempt to equate the two but rather assume that temporal variations in Ct values are associated in changes in infectiousness. To map the two, we used the infectiousness profile estimated from the data of [21], with the correction proposed by [16], *i.e.* a shifted Gamma distribution, with shape 97.19, rate 3.72, and shift 25.63. We then inferred a linear relationship between the Ct value and the instantaneous infectiousness, *i.e.* the infectiousness profile density.We then used this mapping to infer a transmission potential, which can be seen as an infection fitness value [22], using the outputs of the linear mixed model. This was done by transforming Ct values from day 0 into infectiousness values and performing an integration from day 0 to day 21.

References

[1] Davies NG, et al. (2021) Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature* in press.

[2] Volz E, et al. (2021) Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* pp. 1–17.

[3] Challen R, et al. (2021) Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: Matched cohort study. *BMJ* 372:n579.

[4] Faria NR, et al. (2021) Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*.

[5] Li Q, et al. (2021) SARS-CoV-2 501Y.V2 variants lack higher infectivity but do have immune escape. *Cell* 184(9):2362–2371.e9.

[6] Zhou D, et al. (2021) Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* 184(9):2348–2361.e6.

[7] Haim-Boukobza S, et al. (2021) Detection of Rapid SARS-CoV-2 Variant Spread, France, January 26–February 16, 2021. *Emerging Infectious Diseases* 27(5):1496–1499.

[8] Roquebert B, et al. (2021) SARS-CoV-2 variants of concern are associated with lower RT-PCR amplification cycles between January and March 2021 in France. *medRxiv* p. 2021.03.19.21253971.

[9] Canini L, Perelson AS (2014) Viral kinetic modeling: state of the art. *J Pharmacokinet Pharmacodyn* 41(5):431–443. Citation Key Alias: CaniniPerelson2014a.

[10] Néant N, et al. (2021) Modeling SARS-CoV-2 viral kinetics and association with mortality in hospitalized patients from the French COVID cohort. *Proceedings of the National Academy of Sciences* 118(8).

[11] Hellewell J, et al. (2021) Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Medicine* 19(1):106.

[12] Gonçalves A, et al. (2021) SARS-CoV-2 viral dynamics in non-human primates. *PLOS Computational Biology* 17(3):e1008785. Publisher: Public Library of Science.

[13] Sofonea MT, Boennec C, Michalakis Y, Alizon S (2021) Two waves and a high tide: the COVID-19 epidemic in France. *Anaesthesia Critical Care & Pain Medicine* p. 100881.

[14] Alizon S, et al. (2021) Epidemiological and clinical insights from SARS-CoV-2 RT-PCR cycle amplification values. *medRxiv* p. 2021.03.15.21253653.

[15] Hay JA, Kennedy-Shaffer L, Kanjilal S, Lipsitch M, Mina MJ (2020) Estimating epidemiologic dynamics from single cross-sectional viral load distributions. *medRxiv* p. 2020.10.08.20204222.

[16] Ashcroft P, et al. (2020) COVID-19 infectivity profile correction. *Swiss Medical Weekly* 150(3132).

[17] Cosentino G, et al. (2021) SARS-CoV-2 viral dynamics in infections with variants of concern in the French community.

[18] Marks M, et al. (2021) Transmission of COVID-19 in 282 clusters in Catalonia, Spain: A cohort study. *The Lancet Infectious Diseases* 0(0).

[19] Lythgoe KA, et al. (2021) SARS-CoV-2 within-host diversity and transmission. *Science*.

[20] Michalakis Y, Sofonea MT, Alizon S (2021) SARS-CoV-2 viral RNA is not viral load. *OSF Preprints*.

[21] He X, et al. (2020) Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* 26(5):672–675.

[22] Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP (2007) Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc. Natl. Acad. Sci. USA* 104(44):17441–17446.

Authors contribution

MTS, SHB, BR, and SA conceived the project, EL, STP, VF, SHB, and BR collected the data, BE, MTS, and SA analysed the data, JG provided statistical expertise, BE and SA wrote the first draft the manuscript, and all authors contributed to the final version of the manuscript.