Zhao et al. Theoretical Biology and Medical Modelling (2021) 18:10
https://doi.org/10.1186/s12976-021-00140-3

Theoretical Biology and Medical Modelling

RESEARCH Open Access

Modelling the association between COVID-19 transmissibility and D614G substitution in SARS-CoV-2 spike protein: using the surveillance data in California as an example

Shi Zhao[1,2†], Jingzhi Lou[1†], Lirong Cao[1], Hong Zheng[1], Marc K. C. Chong[1,2], Zigui Chen[3], Benny C. Y. Zee[1,2], Paul K. S. Chan[3] and Maggie H. Wang[1,2*]

Abstract

Background: The COVID-19 pandemic poses a serious threat to global health, and pathogenic mutations are a major challenge to disease control. We developed a statistical framework to explore the association between molecular-level mutation activity of SARS-CoV-2 and population-level disease transmissibility of COVID-19.

Methods: We estimated the instantaneous transmissibility of COVID-19 by using the time-varying reproduction number (R_t). The mutation activity in SARS-CoV-2 is quantified empirically depending on (i) the prevalence of emerged amino acid substitutions and (ii) the frequency of these substitutions in the whole sequence. Using the likelihood-based approach, a statistical framework is developed to examine the association between mutation activity and R_t. We adopted the COVID-19 surveillance data in California as an example for demonstration.

Results: We found a significant positive association between population-level COVID-19 transmissibility and the D614G substitution on the SARS-CoV-2 spike protein. We estimate that a per 0.01 increase in the prevalence of glycine (G) on codon 614 is positively associated with a 0.49% (95% CI: 0.39 to 0.59) increase in R_t, which explains 61% of the R_t variation after accounting for the control measures. We remark that the modeling framework can be extended to study other infectious pathogens.

Conclusions: Our findings show a link between the molecular-level mutation activity of SARS-CoV-2 and population-level transmission of COVID-19 to provide further evidence for a positive association between the D614G substitution and R_t. Future studies exploring the mechanism between SARS-CoV-2 mutations and COVID-19 infectivity are warranted.

Keywords: COVID-19, Spike protein, Mutation, Transmission, Statistical modeling

* Correspondence: maggiew@cuhk.edu.hk
†Shi Zhao and Jingzhi Lou contributed equally to this paper, and thus they are joint first authors.
[1]JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China
[2]CUHK Shenzhen Research Institute, Shenzhen, China
Full list of author information is available at the end of the article

Introduction

Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first reported in 2019 [1–5]. The COVID-19 pandemic poses a serious threat to global health and has spread to over 200 countries globally in a short period of time [6, 7]. In response to the ongoing COVID-19 pandemic, the World Health Organization (WHO) declared a public health emergency of international concern on January 30, 2020 [8]. As of September 6, 2020, over 27 million COVID-19 cases have been confirmed worldwide, with over 0.8 million deaths associated with COVID-19 [9].

The dynamics of the transmission of an infectious disease are largely determined by the pathogen's infectiousness and the course of the transmission [10–12]. As a contagious disease with high transmissibility, the control of COVID-19 requires knowledge of the driving factors that may affect disease transmission [13–16]. Pathogenic mutations in SARS-CoV-2 are a major challenge for controlling COVID-19 [17, 18]. Early in February 2020, genetic variants with the D614G substitution on the SARS-CoV-2 spike (S) protein began to spread first in Europe [19] and globally and were suspected to potentially affect viral transmission [20]. Here, 'D614G' denotes the amino acid substitution that changes aspartic acid (D) to glycine (G) on codon 614 of the S protein of SARS-CoV-2. However, the evident relationship between the molecular-level mutation activity of SARS-CoV-2 and the population-level transmissibility of COVID-19 remains unrevealed.

It is biologically reasonable that mutations in viral genomes may alter the pathogenic profile in terms of viral fitness and functionality [21, 22] and consequently change its transmissibility. Previous literature about seasonal influenza epidemics [23] suggested that a few key amino acid substitutions may lead to remarkable changing dynamics of epidemiological outcomes at the population scale. In this study, we adopted a statistical framework to explore and examine the association between COVID-19 transmissibility and key mutation activities in the S protein of SARS-CoV-2.

Data and methods

SARS-CoV-2 sequencing data and COVID-19 surveillance data

The full-length human SARS-CoV-2 strains in California were collected via the Global Initiative on Sharing All Influenza Data (GISAID) [24] on May 24, 2020. A total of 524 strains were searched with collection dates ranging from January 22, 2020, to May 8, 2020. Table 1 summarizes the total number of strains in GISAID and the sample size included in this study for different periods. Since the number of sample stains varied by period, we set 9 successive periods and downloaded a stable number of strains for each period. In the period when more than 30 strains were available, we randomly sampled 30 strains. This sampling scheme is purposely designed to balance the weights due to different sample sizes that may affect the sliding window framework applied in quantifying the mutation activity (details in the next section). Sequences of all SARS-CoV-2 strains acquired are provided in the Additional file 1.

Table 1 Number of human SARS-CoV-2 strains in GISAID and the sample sizes included in this study for different periods

Period	Number of strains	
	in GISAID	in this study
Jan 1–31	6	6
Feb 1–29	16	16
Mar 1–10	72	30
Mar 11–20	94	30
Mar 21–31	158	30
Apr 1–10	100	30
Apr 11–20	21	21
Apr 21–30	50	30
May 1–10	7	7
Total	524	199

Multiple sequence alignment was performed using Clustal Omega (accessed via https://www.ebi.ac.uk/Tools/msa/clustalo/), and the SARS-CoV-2 strain '*China/Wuhan-Hu-1/2019|EPI_ISL_402125*' was considered as the reference sequence. The surveillance data of the daily number of COVID-19 cases in California were collected from the **R** package "*nCov2019*" [25] and The New York Times, accessed via https://github.com/nytimes/covid-19-data and https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html, respectively. Figure 1a shows the daily number of COVID-19 cases in California in a time series.

Instantaneous reproduction number and study period

We adopted the time-varying reproduction number (R_t) to quantify the instantaneous COVID-19 transmissibility in California. Using the framework in [26], we estimated the time-varying reproduction number (R_t) to quantify the instantaneous transmissibility of COVID-19 in California. Following the estimation framework developed in previous studies [26, 27], the epidemic growth of COVID-19 was modeled as a branching process, and thus, R_t can be expressed by using the renewable equation as follows:

$$R(t) = \frac{C(t)}{\int_0^\infty w(k)C(t-k)\mathrm{d}k},$$

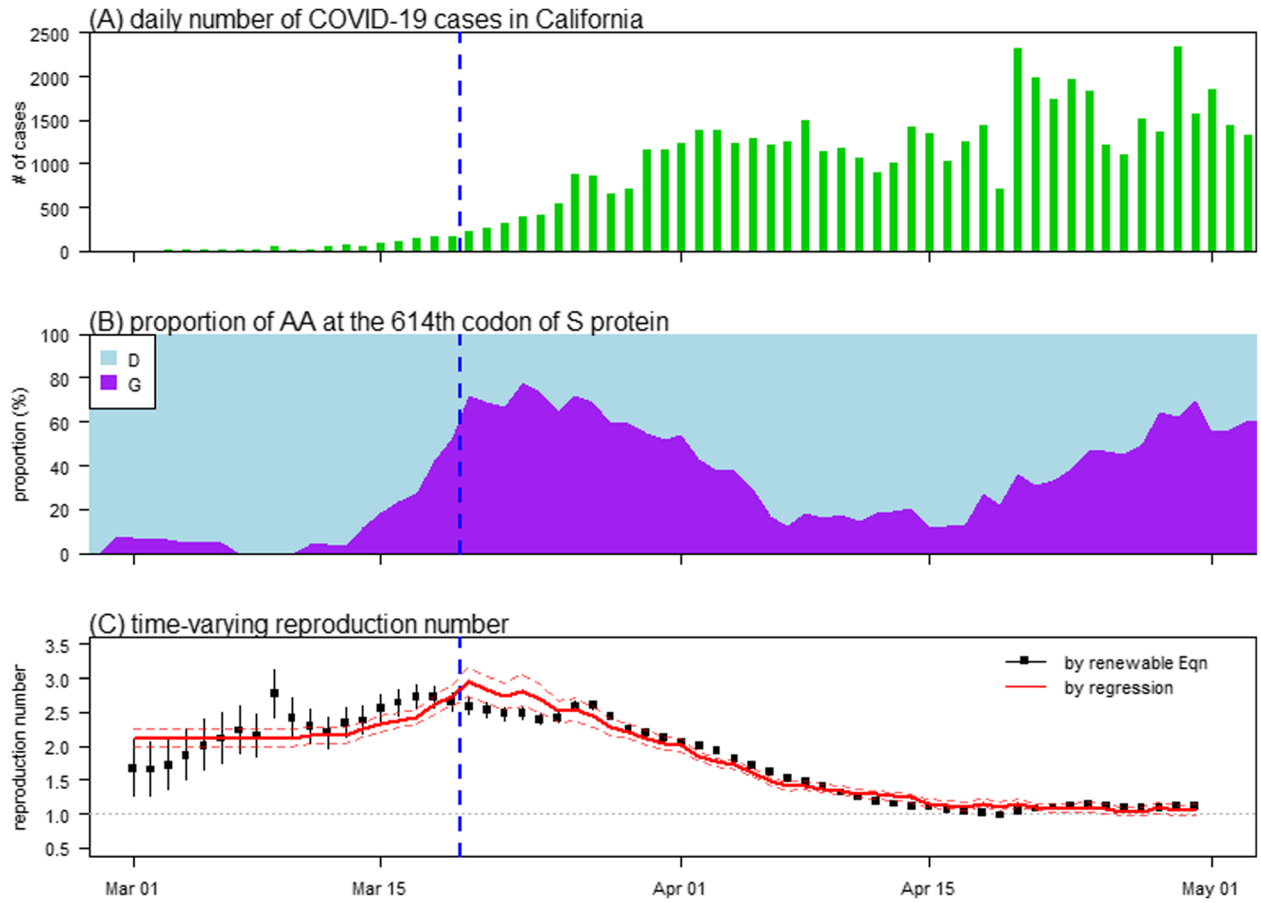

Fig. 1 The number of COVID-19 cases (panel A), prevalence of the amino acids (AA) on the 614-th codon of the S protein (panel B), and time-varying reproduction number (R_t, panel C) in California from March to April 2020. Panel A shows the daily number of COVID-19 time series. Panel B shows the prevalence of the AAs, including Aspartic Acid (D, in cyan) and Glycine (G, in purple), on the 614-th codon of the S protein. Panel C shows the R_t estimated (black) from the number of cases data using renewable equation and fitted (red) by using the mutation activity on the 614-th codon of the S protein in panel B. In each panel, the vertical dashed blue line represents the date, March 19, when the 'stay-at-home' order was officially implemented in California

where $C(t)$ is the number of new COVID-19 cases reported at the t-th date. The function $w(\cdot)$ is the distribution of the generation time (GT) of COVID-19. By averaging the GT estimates from the existing literature [28–35], we considered w as a Gamma distribution with a mean (±SD) value of 5.3 (±2.1) days. Slight variations in the settings of the GT did not affect our main findings.

For the selection of the study period, we considered both the quality of datasets and the increasing intensity (or effects) of local control measures. The selected study period for the COVID-19 surveillance data in California was from March 1, 2020, to April 30, 2020. During this study period, local COVID-19 surveillance was already following the governmental protocol, and the composition of disease control measures was relatively simple and adjustable in further multivariate analyses. In particular, an official 'stay-at-home' order was issued on (t_0 =) March 19, 2020, in California (see https://covid19.ca.gov/stay-home-except-for-essential-needs/), which may affect the patterns of R_t. Hence, we accounted for the effect of this local control measure in further multivariate analyses.

Our analyses depended on both (i) the quality of the data and (ii) the effects of the covariates, especially public health control measures that may decrease R_t. Thus, one of the other reasons, which limited us to consider time outside the study period from March 1, 2020, to April 30, 2020, is related to the prevalence of mutation activities in SARS-CoV-2. During this study period, D614G appears to be the only major amino acid (AA) substitution in the S protein. Thus, complex interactive effects of multiple mutations on infectivity are less likely. As such, our analysis is simplified and is restricted in examining the effect of a single AA substitution.

Quantifying the time-varying molecular-level mutation activity

In previous studies [36–38], a statistical framework was proposed to quantify genetic mutation activities associated with population-level outbreak situations by a metric, namely, the g-measure, on a real-time basis. The g-measure is an empirical time-varying metric calculated from the sequencing data of the pathogen and is determined by a predefined dominance prevalence threshold, θ, ranging from 0 to 1. The θ is the mutation prevalence threshold above which a molecular-level mutation (or substitution) is considered to affect the changing dynamics of the outbreak situation at the population level. The g-measure quantifies the level of key substitutions on a real-time basis, which allows one to explore its linkage to other time-varying variables [39].

We calculated the daily prevalence of amino acids (AA) on each codon in the S protein of SARS-CoV-2. We use $p_{ij}(t)$ to denote the prevalence of the i-th type of amino acid (AA) on the j-th codon of the S protein at time (or date) t, for $i = 1, 2, \ldots, 20$, $j = 1, 2, \ldots, 1273$, and t ranging from January 22 to May 8, 2020. Then, for each AA (20 in total) on each codon (1273 in total) of the S protein, we empirically calculated the prevalence time series. A sliding window was applied to the whole study period, from January 22, 2020, to May 8, 2020, to address the problem of the insufficient daily sample size. Let W denote the window size that represents a constant period (e.g., one week or one month). Hence, for $p_{ij}(t)$ on date t, we accounted for the proportion of the i-th AA out of all 20 types of AAs on the j-th codon within the time period of $t \pm W/2$. In this study, we set W at 7 days for convenience, and we concluded that a variation in W did not affect our main results.

This sliding window scheme requires that the daily sample sizes of sequencing data are close in scale [38]. This guarantees that the prevalence series can reveal the real-world changing patterns of the mutation activity rather than bias towards a particular period with a large number of sequencing samples. Otherwise, as a simple example, during the periods before or after date t, i.e., from $t - W/2$ to t and from

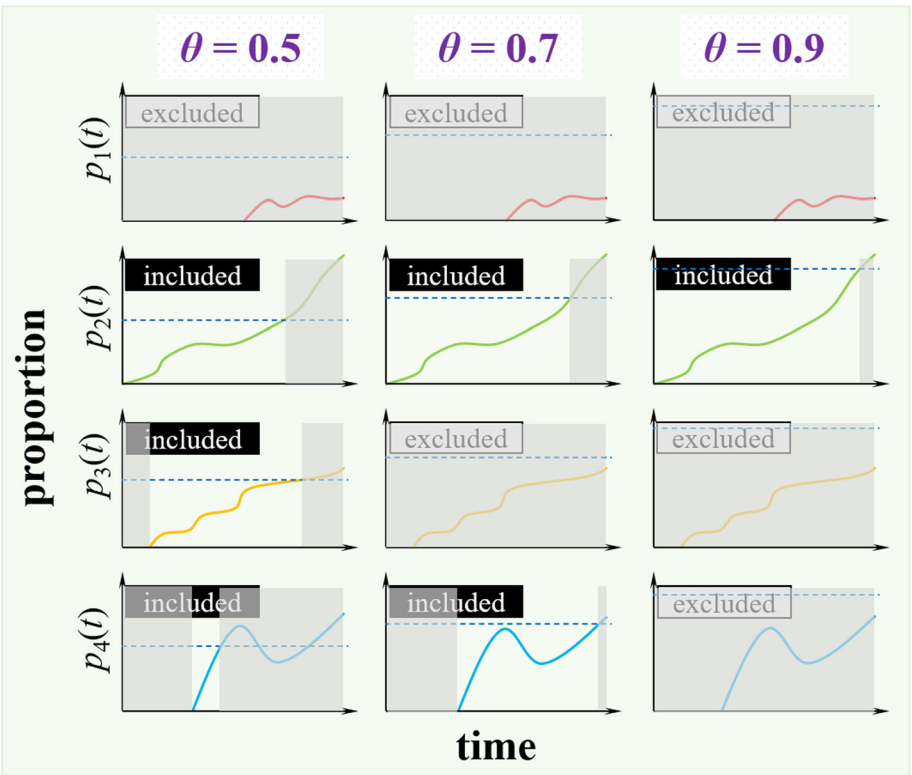

Fig. 2 Illustration diagram of the analytical procedure of g-measure calculation. The prevalence time series of 4 different AA substitutions are denoted by $p_1(t)$, $p_2(t)$, $p_3(t)$ and $p_4(t)$, and indicated in red, green, orange and blue, respectively. Three scenarios of dominant prevalence threshold parameter, θ, are demonstrated with $\theta = 0.5$, 0.7, and 0.9, respectively, which is indicated by the horizontal dashed line in each panel. The g-measure counts the segments of the prevalence series that start from 0 and increase over θ. In each panel, the prevalence series that accounts for g-measure is not shaded in grey region. In other words, the shaded regions are those part of prevalence series excluded from the g-measure calculation. For those prevalence series that never excess θ, they are excluded from g-measure calculation as labeled by 'excluded'; otherwise, 'included' label is indicated

t to $t + W/2$, the prevalence may approach one period with a larger sample size.

Following the calculations from previous studies [36, 37, 39], the g-measure counts the segments of the prevalence series that start from 0 and increase and eventually hits the level of θ. Prevalence series that never excess θ are excluded from the g-measure calculation. For other prevalence series that excess θ at some time point, only those parts start from 0 and increase and hit the level of θ are included in the g-measure calculation. An illustration diagram of the g-measure calculation is presented in Fig. 2. Technically, the algorithm in Table 2 is used to find the indicator function, $I(t)$, to identify the segments of the prevalence series for the g-measure calculation. Therefore, given θ, the g-measure on date t, denoted by g-measure$_t(\theta)$, can be calculated as follows:

$$\text{gmeasure}_t(\theta) = \sum_j \sum_i p_{ijt} \cdot I_{ijt}(\theta) \tag{1}$$

The g-measure quantifies genetic mutation activities and is used to explore the association with R_t. The parameter θ is estimated with the likelihood framework that will be introduced in the remaining parts.

Figure 3 shows the g-measure time series of the S protein with different values of θ. Note that only the g-measure time series from March 1, 2020, to April 30, 2020, were used in further regression analyses.

Regression model and estimation of dominance prevalence threshold

We intended to explore the association between R_t and the mutation activity (measured by the g-measure) on the S protein. A multivariable regression model was fitted to examine the association between R_t and the g-measure considering the effect of local control measures in California.

Since R_t may be affected by disease control measures, we included a dummy variable with a discontinuity design to govern the effect of local control measures. In particular, the official 'stay-at-home' order was issued in California on March 19, 2020 (see https://covid19.ca.gov/stay-home-except-for-essential-needs/#stay-home-order). Hence, in the generalized linear regression model with discontinuity design, we set the structural break in the trends of R_t on March 19, 2020, which was denoted as t_0. In previous studies, R_t is commonly modeled as a Gamma process [26, 40, 41], and thus, the regression is formulated in Eqn (2).

$$\begin{aligned}\mathbf{E}[\ \ln(R_t)] = c + a\,\text{gmeasure}_t \\ + b\,\mathbf{I}(t > t_0)\,(t - t_0)\end{aligned} \tag{2}$$

Here, $\mathbf{E}[\cdot]$ is the function of the expectation. $\mathbf{I}(\cdot)$ is an indicator function that uses the binary variable (0 or 1);if variable t is larger than the threshold value t_0, then 1; otherwise, it is 0. c is the constant parameter, and a and b are the slope parameters. Again, we fixed the term t_0 to be March 19, 2020. The percentage change rate (η) of R_t associated with a 0.01 increase in the g-measure can be calculated directly from the slope parameter a. Thus, the term η is the effect size to be estimated of mutation activity on COVID-19 transmission, and we have $\eta = [\exp(a \times 0.01) - 1] \times 100\%$.

Following previous studies [40, 41], we considered R_t to follow a Gamma process with both means R_t and SDs v_t determined by the renewable equation. For a given time t, the Gamma distribution is denoted by $h(\cdot|R_t, v_t)$, and we model exp. $[c + a \cdot \text{gmeasure}_t + b \cdot \mathbf{I}(t > t_0) \cdot (t - t_0)]$, which is the exponential of the right-hand side of Eqn (2), following the distribution $h(\cdot|R_t, v_t)$. Thus, $h(\cdot|R_t, v_t)$ is a function of parameters a and θ in Eqns (1) and (2), respectively, i.e., $h(a, \theta \mid R_t, v_t)$. In other words, both R_t and v_t were reconstructed directly from the number of cases in a time series (i.e., the raw data) and then served as the known parameters in the likelihood function L, which is given as follows:

$$L(a, \theta) = \prod\nolimits_t h(a, \theta | R_t, v_t).$$

The dominance prevalence threshold parameters θ and a, and equivalently η, can be estimated based on this likelihood framework and the regression model. Then, we calculated the maximum likelihood estimation (MLE) of θ to determine the g-measure for regression analysis. Using the likelihood framework, we estimated the MLE of the dominance prevalence threshold parameter θ, which was adopted to determine the g-measure and to examine the association with R_t. The 95% CIs of the regression parameters were estimated by their point estimates plus or minus Student's t distributed quantile multiplied by their standard errors. Since η and a are one-to-one mappings, the 95% CI of η can also be directly calculated from the 95% CI of a.

We employed Efron's pseudo R-squared and likelihood-based partial R-squared to evaluate the goodness-of-fit of the regression model. A likelihood-ratio (LR) test on the scenarios with (as the full model) and without (as the baseline model) the g-measure was used to examine the reasonability of the model structure.

Table 2 Algorithm of g-measure indicator function, $I(t)$

input: discretized prevalence time series, $p_{1:T}$; dominance prevalence threshold, θ (> 0).
initialization: parameter for recoding the zero-prevalence time point, $\xi = 1$, parameter for recoding excess time point, $\sigma = 0$, $I_{1:T} = \mathbf{0}$.
for t in 1:T **do**
If $p_t == 0$, set $\xi = t$.
If $(p_t \geq \theta \ \&\ \xi > \sigma)$, $I_{(\xi+1):(t-1)} = \mathbf{1}$, $\sigma = t$.
end for
output: discretized indicator time series, $I_{1:T}$.

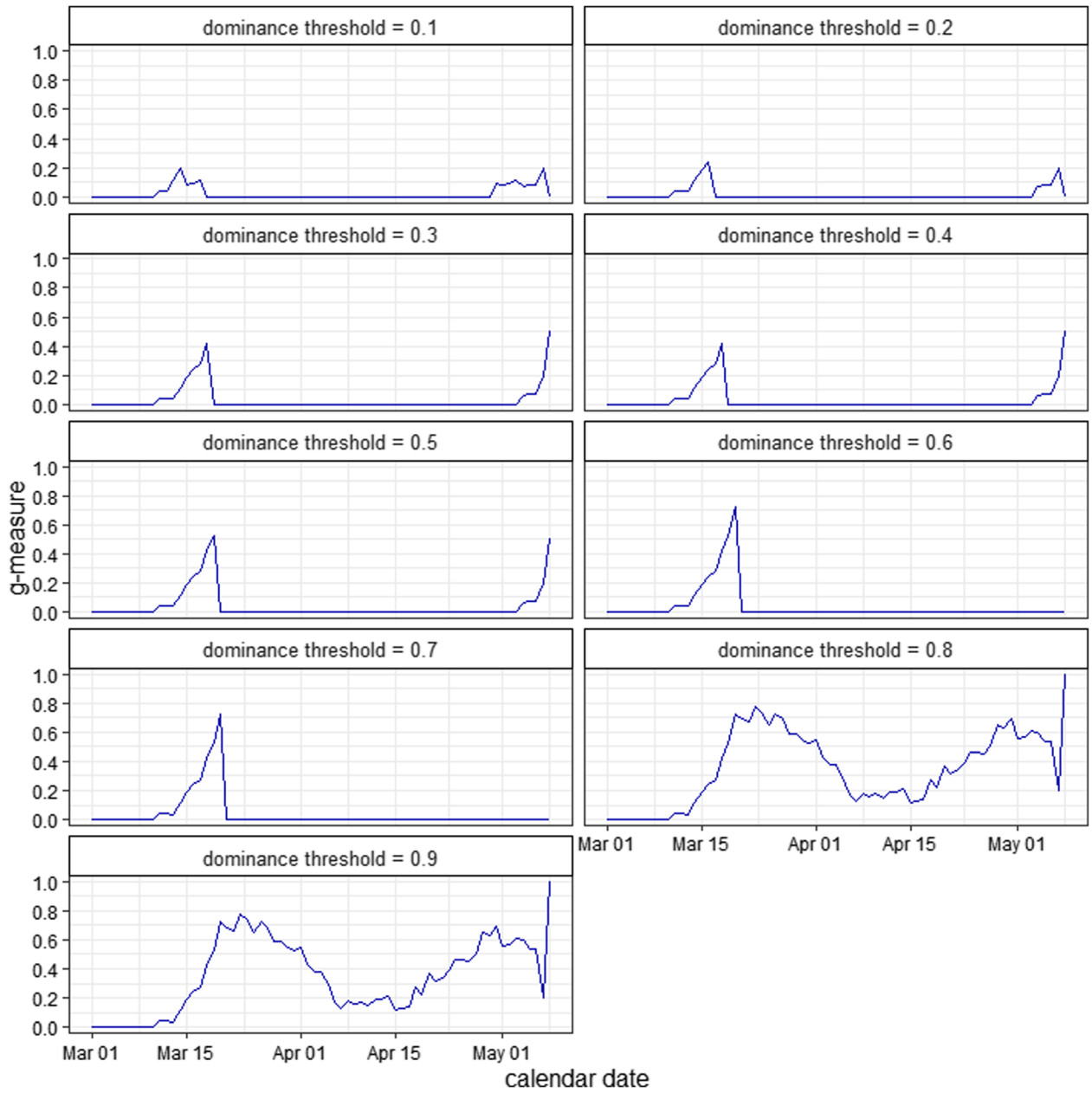

Fig. 3 The g-measure time series of the SARS-CoV-2 spike (S) protein with different values of dominance threshold (θ) using Eqn (1)

Sensitivity analysis

Sensitivity analysis was carried out on the robustness and significance of the association between R_t and mutation activity. We conducted a sensitivity check on the effect size of mutation activity on the S protein in association with the changing dynamics of COVID-19 transmissibility in terms of the reconstructed R_t. We considered three alternative regression formulas, which are similar to the main model Eqn (2), as follows:

$$\mathbf{E}[\ \ln(R_t)] = c + a\,\text{gmeasure}_t \tag{3}$$

$$\mathbf{E}[\ \ln(R_t)] = c + a\,\text{gmeasure}_t + b\,\mathbf{I}(t > t_0) \tag{4}$$

$$\begin{aligned}\mathbf{E}[\ \ln(R_t)] = {} & c + a\,\text{gmeasure}_t \\ & + b\,\mathbf{I}(t > t_0)\,(t - t_0) \\ & + d\,\mathbf{I}(t \le t_0)\,(t - t_0)\end{aligned} \tag{5}$$

To check the robustness and significance of the estimates, we examined the consistency of both the sign (i.e., + or −) and the statistical significance (in terms of p-value < 0.05) of the regression coefficient a in the four models in Eq. (2)–(5).

Results and discussion

We reconstructed the daily instantaneous reproduction number (R_t) from the epidemic curve, as shown in Fig. 1c (black dots). We observed that the overall trends of R_t were relatively steady in the first half of March but gradually decrease thereafter since the local 'stay-at-home' order was issued in California on March 19, 2020, which was adjusted in Eqn (2). During the first half of March, which was regarded as the early phase of the outbreak, the reproduction number ranged from 1.5 to 3, and this range is generally consistent with previous estimates [2, 3, 6, 12, 29, 33, 42–48].

We estimated the dominance prevalence threshold (θ) at 0.8, as shown in Fig. 4, which was adopted to examine the association between the g-measure and R_t. When $\theta = 0.8$, we found that the g-measure of the S protein appeared to be solely contributed to by the D614G substitution (see Fig. 1b), which also holds for all θ values > 0.75. In other words, the D614G substitution is considered a key mutation and is likely dominant in accounting for the changes in COVID-19 transmissibility due to a mutation at the molecular level.

Using the regression model in Eqn (2), we found a significant positive association between the g-measure and R_t when $\theta = 0.8$ (as estimated). Hence, the changing dynamics of R_t are likely associated with the key mutations that are solely contributed to by the D614G substitution. We estimated that each 0.01 increase in the prevalence of glycine (G) on codon 614 is positively associated with a 0.49% (95% CI: 0.39 to 0.59) increase in R_t, which, in terms of the partial R-squared, explains 61% of the R_t variation after accounting for the control measures. Figure 1c shows the fitting results by using the regression model in Eqn (2). By examining the patterns in Fig. 1, we found that the prevalence of the D614G substitution matches the trends of R_t in March 2020. However, we noticed that since (roughly) April 15, 2020, the prevalence of the D614G substitution increased, but R_t remained constant. The reasons may include that the increase in transmissibility was counteracted by the effects of local nonpharmaceutical interventions that reduced the transmission of COVID-19. Sensitivity analysis with alternative model structures in Eqns (3)–(5) indicates that the positive association between the D614G substitution and R_t holds robustly and significantly (data not shown).

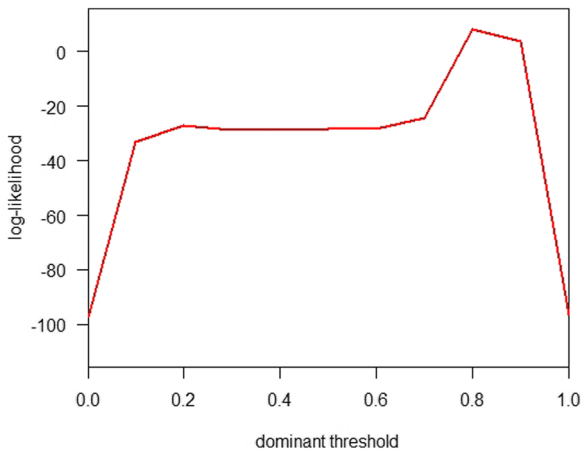

Fig. 4 The likelihood profile of the dominant prevalence threshold parameter, θ, using the likelihood framework associated with regression model in Eqn (2)

The significant positive association between the D614G substitution and R_t is biologically reasonable and consistent with findings in previous studies. The few (but key) AA substitutions may vary the three-dimensional structure of the protein as well as influence the receptor binding process in which a pathogen invades host cells. Previous analysis implied that the D614G substitution may alter the conformation of the S protein and thus may theoretically functionally enhance receptor binding capacity [19, 20, 49, 50], leading to an increase in SARS-CoV-2 transmissibility and pathogenicity [51]. Similarly, we learn from the influenza virus that major antigenic changes can be caused by a single AA substitution related to the receptor binding domain (RBD) [52]. Our analytical framework is data-driven and can be extended to study other infectious diseases.

For the limitations of this study, we have the following remarks. First, the reconstruction of R_t relies on the setting of the generation time (GT). We modeled the distribution of COVID-19 GT as a fixed Gamma distribution, which follows previous findings [28–32]. In a real-world situation, the time interval between the transmission generations could be variable [42, 53], which may affect the estimation of R_t. However, the changes in R_t estimates due to slight variations in GT are negligible [42]. We remark that this issue is unlikely to affect our main conclusion, and our model can be extended to a more complex context with the available time-varying GT data. Second, for the R_t estimation parts, $C(t)$ should be the number of COVID-19 cases onset at time t. However, because the data by onset are unavailable, we adopted the current dataset by reporting data as a proxy for the COVID-19 incidence time series. If one considers a constant reporting lag, the R_t estimates will be in exactly the same trends but shifted for the reporting lag. Considering that a similar reporting delay also occurred for the SARS-CoV-2 sequencing data, the effects of the two reporting lags may be counteracted. We note that this approximation in our analysis is unlikely to affect the conclusion in this study. In addition, with detailed reporting lag information of each individual case, adjustment for the reporting delay can surely be carried out based on our current analytical framework. Third, as a

data-driven study, the estimated association should be interpreted with caution. With ecological settings, our analysis provides statistical evidence about the likelihood of causality, but the findings in this study cannot guarantee causality, which needs further biomedical experiments with a more sophisticated context.

Conclusions

Our findings show a link between the molecular-level mutation activity of SARS-CoV-2 and population-level COVID-19 transmission to provide further evidence for a positive association between the D614G substitution and R_t. Future studies exploring the mechanism between SARS-CoV-2 mutations and COVID-19 infectivity are warranted.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12976-021-00140-3.

Additional file 1:.

Abbreviations

AA: Amino acid; COVID-19: Coronavirus disease 2019; D614G: amino acid substitution of aspartic acid (D) to glycine (G) on codon 614 (of the S protein of SARS-CoV-2); GT: Generation time; LR: Likelihood ratio; GISAID: Global Initiative on Sharing All Influenza Data; MLE: Maximum likelihood estimation; RBD: Receptor binding domain; SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2; SD: Standard deviation; 95% CI: 95% confidence interval

Acknowledgments

This study is conducted using the resources of Alibaba Cloud Intelligence High Performance Cluster computing facilities, which is made free for COVID-19 research.

Disclaimer

The funding agencies had no role in the design and implementation of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Authors' contributions

SZ and MHW conceived the study. JZ collected the data. SZ and JL carried out the analysis and drafted the first manuscript. SZ, JL and MHW discussed the results. All authors critically read and revised the manuscript and gave final approval for publication.

Funding

This work is supported by CUHK grants [PIEF/Ph2/COVID/06, 4054456] and the Health and Medical Research Fund (HMRF) Commissioned Research on COVID-19 [INF-CUHK-1] of Hong Kong, China and is partially supported by the National Natural Science Foundation of China (NSFC) [31871340,71974165].

Availability of data and materials

All data used in this work are publicly available.

Ethics approval and consent to participate

The number of COVID-19 cases and sequencing data are collected via public domains, and thus, neither ethical approval nor individual consent is applicable.

Consent for publication

Not applicable.

Competing interests

MHW is a shareholder of Beth Bioinformatics Co., Ltd. BCYZ is a shareholder of Beth Bioinformatics Co., Ltd. and Health View Bioanalytics, Ltd. The other authors declare no competing interests.

Author details

[1]JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China. [2]CUHK Shenzhen Research Institute, Shenzhen, China. [3]Department of Microbiology, Chinese University of Hong Kong, Hong Kong, China.

Received: 8 September 2020 Accepted: 12 February 2021
Published online: 09 March 2021

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England). 2020;395(10223):497–506.
2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382(13):1199–207.
3. Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven Modelling analysis of the early outbreak. J Clin Med. 2020;9(2):388.
4. Leung K, Wu JT, Liu D, Leung GM. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. Lancet (London, England). 2020;395(10233):1382–93.
5. Parry J. China coronavirus: cases surge as official admits human to human transmission. BMJ (Clin Res ed). 2020;368:m236.
6. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet (London, England). 2020; 395(10225):689–97.
7. Zhao S, Zhuang Z, Cao P, Ran J, Gao D, Lou Y, et al. Quantifying the association between domestic travel and the exportation of novel coronavirus (2019-nCoV) cases from Wuhan, China in 2020: a correlational analysis. J Travel Med. 2020;27(2):taaa022.
8. World Health Organization. Statement on the second meeting of the International Health Regulations Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV), World Health Organization (WHO). 2020 [Available from: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).
9. World Health Organization. Novel Coronavirus (2019-nCoV) situation reports, released by the World Health Organization (WHO). 2020 [Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.
10. Tuite AR, Fisman DN. Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic. Ann Intern Med. 2020;172(8):567–8.
11. Zhao S, Cao P, Gao D, Zhuang Z, Cai Y, Ran J, et al. Serial interval in determining the estimation of reproduction number of the novel coronavirus disease (COVID-19) during the early outbreak. J Travel Med. 2020;27(3):taaa033.
12. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro Surv. 2020;25(4):2000058.
13. Zhao S. To avoid the noncausal association between environmental factor and COVID-19 when using aggregated data: simulation-based counterexamples for demonstration. Sci Total Environ. 2020:141590.
14. Kutter JS, Spronken MI, Fraaij PL, Fouchier RA, Herfst S. Transmission routes of respiratory viruses among humans. Curr Opin Virol. 2018;28:142–51.
15. Lau H, Khosrawipour V, Kocbach P, Mikolajczyk A, Schubert J, Bania J, et al. The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. J Travel Med. 2020;27(3):taaa037.
16. Zhao S, Musa SS, Hebert JT, Cao P, Ran J, Meng J, et al. Modelling the effective reproduction number of vector-borne diseases: the yellow fever outbreak in Luanda, Angola 2015-2016 as an example. PeerJ. 2020;8:e8601.

17. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science. 2020;369(6506):1014–8.
18. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020;83:104351.
19. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020;94(7).
20. Benvenuto D, Demir AB, Giovanetti M, Bianchi M, Angeletti S, Pascarella S, et al. Evidence for mutations in SARS-CoV-2 Italian isolates potentially affecting virus transmission. J Med Virol. 2020.
21. Rimmelzwaan GF, Berkhoff EGM, Nieuwkoop NJ, Fouchier RAM, Osterhaus A. Functional compensation of a detrimental amino acid substitution in a cytotoxic-T-lymphocyte epitope of influenza a viruses by comutations. J Virol. 2004;78(16):8946–9.
22. Rimmelzwaan GF, Berkhoff EGM, Nieuwkoop NJ, Smith DJ, Fouchier RAM, Osterhaus A. Full restoration of viral fitness by multiple compensatory co-mutations in the nucleoprotein of influenza a virus cytotoxic T-lymphocyte escape mutants. J Gen Virol. 2005;86(6):1801–5.
23. Gog JR, Rimmelzwaan GF, Osterhaus ADME, Grenfell BT. Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza a. Proc Natl Acad Sci. 2003;100(19):11143–7.
24. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data–from vision to reality. Euro Surv. 2017;22(13):30494.
25. Wu T, Ge X, Yu G, Hu E. Open-source analytics tools for studying the COVID-19 coronavirus outbreak. medRxiv. 2020:2020.02.25.20027433.
26. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505–12.
27. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004; 160(6):509–16.
28. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dorner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science. 2020;368(6491):eabb6936.
29. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. 2020;26(5):672-5.
30. Zhao S. Estimating the time interval between transmission generations when negative values occur in the serial interval data: using COVID-19 as an example. Math Biosci Eng. 2020;17(4):3512–9.
31. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Euro Surv. 2020;25(17):2000257.
32. Tindale LC, Stockdale JE, Coombe M, Garlock ES, Lau WYV, Saraswat M, et al. Evidence for transmission of COVID-19 prior to symptom onset. Elife. 2020;9: e57149.
33. Zhao S, Gao D, Zhuang Z, Chong MKC, Cai Y, Ran J, et al. Estimating the serial interval of the novel coronavirus disease (COVID-19): a statistical analysis using the public data in Hong Kong from January 16 to February 15, 2020. Front Phys. 2020;8:347.
34. Wang K, Zhao S, Liao Y, Zhao T, Wang X, Zhang X, et al. Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China, from 19 January to 2020. Transbound Emerg Dis. 2020;67(6):2818-22
35. Ma S, Zhang J, Zeng M, Yun Q, Guo W, Zheng Y, et al. Epidemiological parameters of coronavirus disease 2019: a case series study. J Med Internet Res. 2020;22(10):e19994.
36. Wang MH, Lou J, Cao L, Zhao S, Chan PKS, Chan MC-W, et al. Characterization of the evolutionary dynamics of influenza A H3N2 hemagglutinin. bioRxiv. 2020:2020.06.16.155994.
37. Wang MH, Lou J, Zee BCY, Chong KC. US Provisional Patent No. 62/687645, 2019 PCT/CN2019/091652. Measurement and Prediction on Influenza Virus Genetic Mutation Patterns. . US Patent. 2018.
38. Zhao S, Lou J, Cao L, Chen Z, Chan RW, Chong MK, et al. Quantifying the importance of the key sites on haemagglutinin in determining the selection advantage of influenza virus: using a/H3N2 as an example. J Inf Secur. 2020; 81(3):452–82.
39. Lou J, Zhao S, Cao L, Chong MK, Chan RW, Chan PK, et al. Predicting the dominant influenza A serotype by quantifying mutation activities. Int J Infect Dis. 2020;100:255-7.
40. Ali ST, Kadi A, Ferguson NM. Transmission dynamics of the 2009 influenza a (H1N1) pandemic in India: the impact of holiday-related school closure. Epidemics. 2013;5(4):157–63.
41. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355–9.
42. Ali ST, Wang L, Lau EHY, Xu XK, Du ZW, Wu Y, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 2020;369(6507):1106–9.
43. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395–400.
44. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, et al. Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. Proc Natl Acad Sci U S A. 2020;117(19):10484–91.
45. Jung SM, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. J Clin Med. 2020;9(2):523.
46. Musa SS, Zhao S, Wang MH, Habib AG, Mustapha UT, He D. Estimation of exponential growth rate and basic reproduction number of the coronavirus disease 2019 (COVID-19) in Africa. Infect Dis Poverty. 2020;9(1):96.
47. Ran J, Zhao S, Han L, Liao G, Wang K, Wang MH, et al. A re-analysis in exploring the association between temperature and COVID-19 transmissibility: an ecological study with 154 Chinese cities. Eur Respir J. 2020;56(2):2001253.
48. Xu XK, Liu XF, Wu Y, Ali ST, Du Z, Bosetti P, et al. Reconstruction of transmission pairs for novel coronavirus disease 2019 (COVID-19) in mainland China: estimation of super-spreading events, serial interval, and hazard of infection. Clin Infect Dis. 2020;71(12):3163–7.
49. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020;182(4):812–27.
50. Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell. 2020;183(3):739–51.
51. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell. 2020;184(1):64-75.
52. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science. 2013;342(6161): 976–9.
53. Zhao S, Cao P, Chong MKC, Gao D, Lou Y, Ran J, et al. COVID-19 and gender-specific difference: analysis of public surveillance data in Hong Kong and Shenzhen, China, from January 10 to February 15, 2020. Infect Control Hosp Epidemiol. 2020;41(6):750–1.

Publisher's Note