GMR

The Original by FUNPEC-RP

Temporal data series and logistic models reveal the dynamics of SARS-CoV-2 spike protein D614G variant in the COVID-19 pandemic

F. Antoneli[1], T.N. Furuyama[3], I.M.V.G. Carvalho[2], M.R.S. Briones[1] and L.M.R. Janini[3]

[1] Centro de Bioinformática Médica, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, Brasil
[2] Laboratório de Parasitologia, Instituto Butantan, São Paulo, Brasil
[3] Laboratório de Retrovirologia e Departamento de Medicina, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo Brasil

Corresponding author: F. Antoneli
E-mail: fernando.antoneli@unifesp.br

Genet. Mol. Res. 20 (4): gmr18960
Received August 24, 2021
Accepted October 27, 2021
Published December 13, 2021
DOI http://dx.doi.org/10.4238/gmr18960

ABSTRACT. The COVID-19 pandemic is caused by the worldwide spread of the RNA virus SARS-CoV-2. Because of its mutational rate, wide geographical distribution, and host response variance this coronavirus is currently evolving into an array of strains with increasing genetic diversity. Most variants apparently have neutral effects for disease spread and symptom severity. However, in the viral spike protein, which is responsible for host cell attachment and invasion, the D614G variant, containing the amino acid substitution D to G in position 614, was suggested to increase viral infection capability. Here we propose a novel method to test the epidemiological impact of emergence of a new variant, by a combination of epidemiological curves (for new cases) and the temporal variation of relative frequencies of the variants through a logistic regression model. We applied our method to temporal distributions of SARS-CoV-2 D614 or G614, in two geographic areas: USA (East Coast versus West Coast) and Europe-Asia (East Countries versus West Countries). Our analysis shows that the D614G prevalence and the growth rates of COVID-19 epidemic data curves are correlated at the early stages and not correlated at the late stages, in both the USA and

Europe-Asia scenarios. These results show that logistic models can reveal the potential selective advantage of D614G, which can explain, at least in part, the impact of this variant on COVID-19 epidemiology.

Key words: COVID-19; SARS-CoV-2; D614G; Mutation; Population genetics; Coronavirus

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is caused by SARS-CoV-2 and was declared a pandemic by the World Health Organization (WHO) on March 11, 2020 (World Health Organization, 2020a; 2020b). As of October 20, 2021, the number of confirmed COVID-19 cases reached 241 million, with 4.9 million deaths worldwide (https://covid19.who.int/). This is the third outbreak caused by a coronavirus in less than 20 years (World Health Organization, 2020a). From November 2002 to May 2004, SARS-CoV-1 (Severe Acute Respiratory Syndrome caused by Coronavirus type 1) affected 26 countries worldwide, accounting for 8,096 confirmed cases and 774 deaths (9.6% fatality ratio) (Drosten et al., 2003; Ksiazek et al., 2003; Lee et al., 2003; Peiris et al., 2003; Zhong et al., 2003; Centers for Disease Control and Prevention - Department of Health and Human Services, 2004; World Health Organization, 2004; Centers for Disease Control and Prevention, 2017). MERS-CoV (Middle East Respiratory Syndrome caused by Coronavirus) spread to 27 countries around the globe, totaling 2,519 confirmed cases and 866 deaths (34.4% fatality ratio) continuously since April 2012 (Zaki et al. 2012; Hijawi et al. 2013; Centers for Disease Control and Prevention 2019; World Health Organization 2019; World Health Organization 2020d). Several conditions contribute to the transmission speed of SARS-CoV-2, such as transmission during the asymptomatic phase and wide human susceptibility to this pathogen (Arons et al., 2020; Fam et al., 2020; Gandhi et al., 2020). The central concern for governments and the general population is the collapse of healthcare systems and lack of essential care. Therefore, more information about SARS-CoV-2 mechanisms inside the host cell, its epidemiology and its genetic patterns are necessary to halt virus spread, to prevent the disease and heal infected individuals.

A comparison of several SARS-CoV-2 strains with the Wuhan reference genome (GenBank accession NC_045512) reveals a G to A transition at position 23,403 that leads to a D to G amino acid substitution at position 614 in the spike protein. Molecular evidence suggests that this substitution is advantageous for viral propagation *in vitro* because of increased spike protein abundance and reduced shedding (Zhang et al., 2020). A previous study by Korber and collaborators (Korber et al., 2020) conjectures that the D614G substitution in SARS-CoV-2 Spike protein could be responsible for higher transmission rates observed on a global scale. The study shows that there was a higher prevalence of D614 in China and in the United States before March 2020, while after March 2020 the G614 prevalence significantly increased in Europe and the United States.

If there is a correlation between the D614G variant prevalence and higher SARS-CoV-2 transmission rate, then the epidemiological data might reveal a significant correlation between D614G prevalence and the growth rate coefficients of epidemic curves globally. Here we propose a method to examine if the prevalence of a variant is correlated with increased growth rate coefficients in temporal series of COVID-19 epidemiological data. Our method differs from other approaches to estimate transmission rates and selective

advantage in that it does not rely on extensive genomic sequencing (Dorp et al., 2020; Lythgoe et al., 2020; Trucchi et al., 2021; Zhang et al., 2020).

MATERIAL AND METHODS

D614G variant data

Data on prevalence of D614 or G614 was obtained from the Los Alamos distribution map of D614 and G614 SARS-CoV-2 (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Los Alamos National Laboratory, 2020). The data was downloaded as "data-2020-05-20.csv" file for "all" time range. The time range considered was 21 days (three weeks). Data: Obtained from Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University; the Red Cross; the Census American Community Survey; the Bureau of Labor and Statistics: (https://github.com/CSSEGISandData/COVID-19).

Epidemic data

The epidemic growth rates were obtained from Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University; the Red Cross; the Census American Community Survey; and the Bureau of Labor and Statistics data (2020a; 2020b; 2020c; Johns Hopkins University 2020). The time range considered was the same as the considered in the Los Alamos data, being the first day of time series of confirmed cases on February 18, 2020, and the last day of time series of confirmed cases on May 27, 2020.

Countries, States, Counties and time ranges

Six states and 325 counties of the USA were considered in the analysis further grouped as East or the West Coast. Accordingly, West Coast States data included (with number of counties in parentheses): Oregon (34), Washington (39), and California (57), with a total of 129 from the West Coast. The East Coast States included (number of counties in parentheses): New York (58), Connecticut (9), and Virginia (130), totaling 196 counties from the West Coast. The time range was divided into an early start period and a late start period. The early start considered counties such that the epidemics started from 2/18/20 – 3/18/20 (98 counties). The late start considered counties which epidemics started from 3/18/20 – 5/27/20 (325 counties). Nine counties did not have at least 21 days of non-zero time series. The analysis was also made for other regions of the world, considering a broader division: World (some countries have more than one administrative region): Western Countries (Europe): Belgium (1), Denmark (3), France (9), Germany (1), Italy (1), Luxembourg (1), Netherlands (5), Portugal (1), Spain (1), United Kingdom (11); total number of administrative regions = 34. Eastern Countries (Eastern Asia and Oceania): Australia (8), Bangladesh (1), China (32), India (1), Japan (1), South Korea (1), Singapore (1), Taiwan (1), Thailand (1), and Vietnam (1); total number of administrative regions = 48. The first day of time series of confirmed cases was 1/22/20 (eastern countries); 2/15/20 (western countries) and the last day of time series of confirmed cases was 5/27/20.

Regarding the early and late epidemic periods, the time frame was: Early start: countries in which the epidemics started from 2/15/20 – 3/15/20 (western countries; 19 regions) and 1/22/20 – 5/27/20 (eastern countries; 47 regions). Late start: countries in which the epidemics started from 3/15/20 – 5/27/20 (western countries; 15 regions) and 1/22/20 – 5/27/20 (eastern countries; 47 regions). One region did not have at least 21 days of non-zero time series and the eastern regions were all early starters, so in this case we compared the eastern regions with the early western region and the late western regions.

Logistic models

The logistic model parameters were obtained from logistic regression (Spiegelhalter 1986) using Python 3 with Pandas libraries (https://pandas.pydata.org/) and scikit-learn (https://scikit-learn.org/stable/about.html). Plots were generated with matplotlub (https://matplotlib.org/) and seaborn (https://seaborn.pydata.org/index.html).

RESULTS

The initial analysis consisted in the logistic models derived from US data comparing US East coast (predominantly G614) with West coast (predominantly D614), the Asia-Europe data in West (predominantly G614) and East (predominantly D614) in the early and late epidemic stages. Plots of logistic models show the corresponding logistic model (blue line) and its confidence band (light blue shading) (Figure 1).

The region-specific data consist of the Early Start Counties of USA (Figure 1A), the Late Start counties of USA (Figure 1B), the Early Start countries of the Asia-Europe axis (Figure 1C) and the Late Start countries of Asia-Europe axis (Figure 1D). The comparisons between the logistic models in early and late epidemic stages show that at the early stages the growth rates between West and East, either US or Europe, are significantly different, while in late stages the West-East differences are not significant. This test therefore suggests that in the early epidemic stages the predominant variant pattern reflected a "founder effect", especially in the US, where the West coast infections derived from an Asian D614 type whereas in the East cost the infection dynamics started with European derived ancestors, of the G614 type. This is observed from D614G distribution data. The early epidemic stage in both US and Asia-Europe show significant differences in the odds ratios in the West and East portions, showing that the growth rates might be impacted by the G614 substitution. At the late stages the growth rates are not distinguishable. European late stages show odds ratios <1 and in the US the odds ratios drop from 1.16 to 1.03.

The growth curves of the D614 and G614 variants in West US, East US, West Asia-Europe and East Asia-Europe at different time periods, of 10 days each, reveal the dynamics of variant D614G (Figure 2). The time series of frequency variants (D/G) reveal that irrespective of geographic region and early and late epidemic stages, the G614 variant increases and surpasses the D614. The dynamics of D614 and G614 in USA (Figure 2A) and Asia-Europe (Figure 2B) show a steep increase of G614 variant. The frequencies of variants in the same region are complementary to each other. In US East the G614 started at approximately 80% and D614 at ~20% with subsequent increase of G614 to 100% and extinction of D614. In US West, G614 started at ~1% and D614 ~99% and after 70 days G614 increased to more than 60% whereas D614 decreased to less than 40% (Figure 2A).

This type of dynamics is highly suggestive of a selective sweep because of an increased infectivity/replication rate of G614. The effect is very similar in Europe West and East where irrespective of the initial frequencies of D614 and G614, the later always predominates after 60 days (Figure 2B).

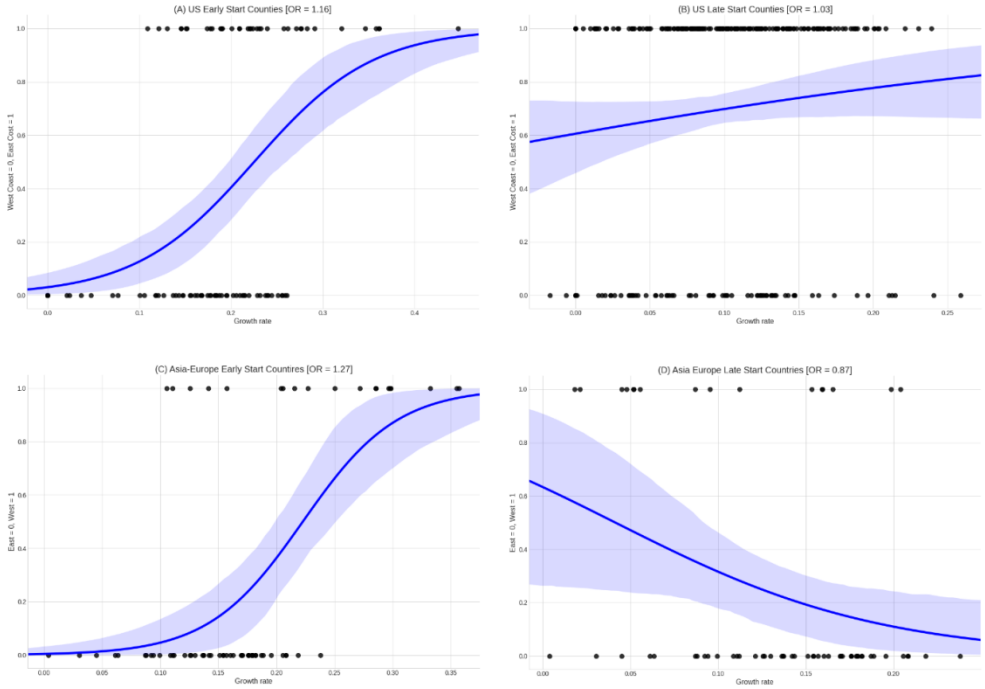

Figure 1. Plots of logistic models. Panel (A) shows the Early Start Counties of USA, panel (B) shows the Late Start counties of USA, panel (C) shows the Early Start countries of the Asia-Europe axis and panel (D) shows the Late Start countries of Asia-Europe axis (black dots). The blue curve is the corresponding logistic model with its confidence band (light blue shading). In each panel, the horizontal axis is the growth rate of the initial segment of 21 days of the corresponding time series of confirmed cases, and the vertical axis is a binary variable indicating the corresponding region where the time series is from (East/west coast counties in the USA, panels (A) and (B) or East/West countries in Asia-Europe axis, panels (C) and (D)).

The frequencies of D614 and G614 are compared in the East and West coast of US (Table 1) while the Asia-Europe frequencies of D614 and G614 are in Table 2. These data show that the growth rate of G614 is significantly higher than the D614 growth rate. Also, it indicates that the phenomenon is global, not restricted to a geographic location of specific host population. The initial cases in the East coast are likely to have originated from European strains (predominantly G614) whereas in the West coast the initial infections were caused by Asian strains D614 predominant in that Continent at that time. The quality of the linear model fitting of the epidemiological data was determined by goodness-of-fit (Figures 3-6). The growth rates were estimated by linear regression of log-transformed data giving the exponential growth of the beginning curve of infected people for each county (US) or country (Asia-Europe).

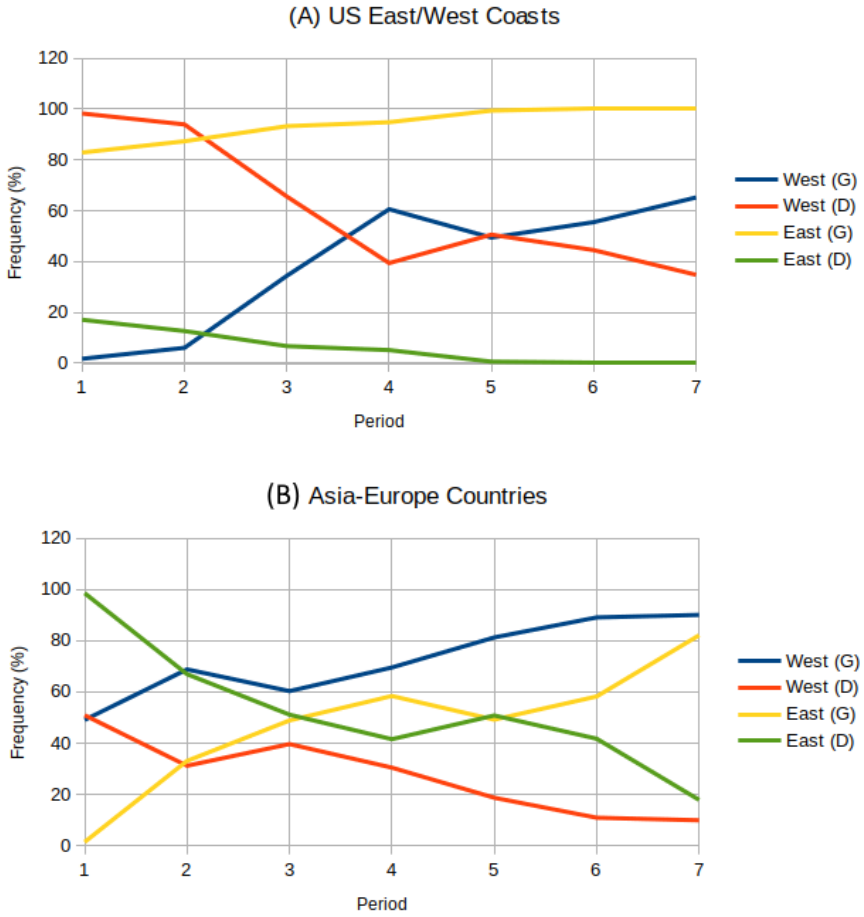

Figure 2. Time evolution of the frequency viral variants (D/G) with respect to the residue 614 of spike protein. Panel **(A)** shows the USA and panel **(B)** shows the Asia-Europe axis. In each panel, the horizontal axis shows time period of time (corresponding to 10 days each) and the vertical axis show the frequency. Each curve corresponds to the variant frequency in the corresponding region, according to the side legends. The frequencies of variants in the same region are complementary to each other. Adapted from (https://cov.lanl.gov/apps/covid-19/map/) (Korber et al. 2020).

Table 1. The frequencies of SARS-Cov2 D614 and G614 variants in the East and West coasts of the US.

US Data						
Period	**West Coast**			**East Coast**		
	G614	**D614**	**Total**	**G614**	**D614**	**Total**
1	1	56	57	0	0	0
2	12	186	198	63	13	76
3	183	349	532	537	78	615
4	320	208	528	304	22	326
5	211	216	427	402	22	424
6	111	89	200	156	1	157
7	60	32	92	41	0	41

Data on the prevalence of variant of the virus with respect to the residue 614 of spike protein is prevalent in the infected population (D or G). Data from the site "Distribution of D614 and G614" (https://cov.lanl.gov/apps/covid-19/map/) from (Korber et al. 2020).

Table 2. The frequencies of SARS-Cov2 D614 and G614 variants in the East and West in the worldwide data.

Worldwide Data						
Period	**West**			**East**		
	G614	**D614**	**Total**	**G614**	**D614**	**Total**
1	53	55	108	11	767	778
2	574	260	834	41	83	124
3	1152	757	1909	259	271	530
4	2986	1299	4267	353	251	604
5	4038	929	4967	151	156	307
6	1946	238	2184	67	48	115
7	892	98	900	69	15	84

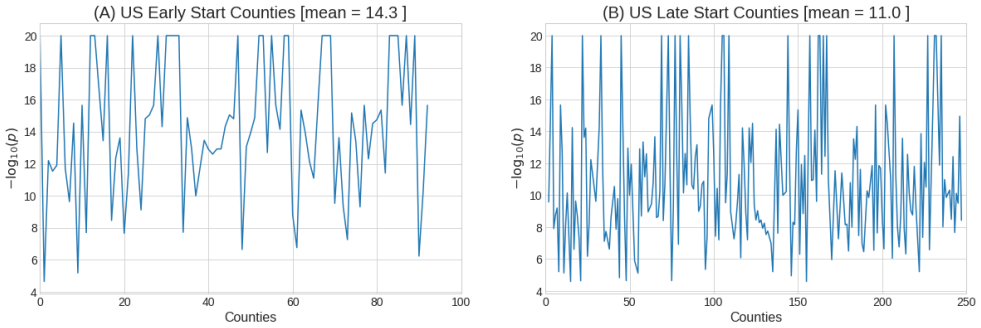

Figure 3. Goodness-of-fit assessment of the linear model of epidemiological data for the US counties. Vertical axis is the $-\log_{10}(p)$, where p is the p-value for the slope given by linear regression of log-transformed data (that is the exponential growth of the beginning curve of infected people). The counties are on the horizontal axis in alphabetical order.

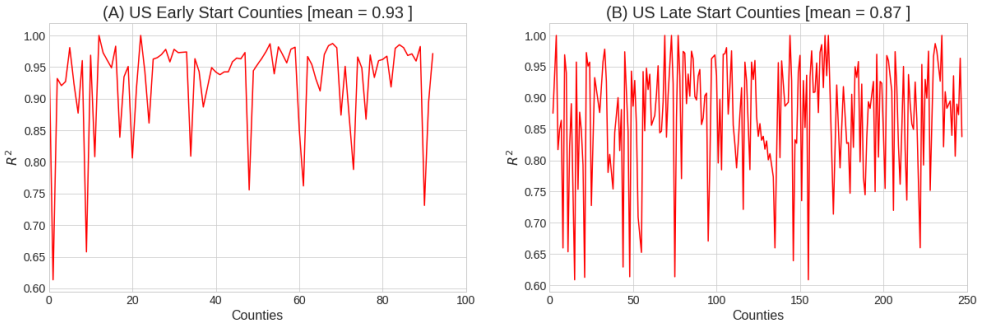

Figure 4. Goodness-of-fit assessment of the linear model of epidemiological data for the US counties. Vertical axis is the R^2 score of the linear regression of log-transformed data (that is the exponential growth of the beginning curve of infected people). The counties are on the horizontal axis in alphabetical order.

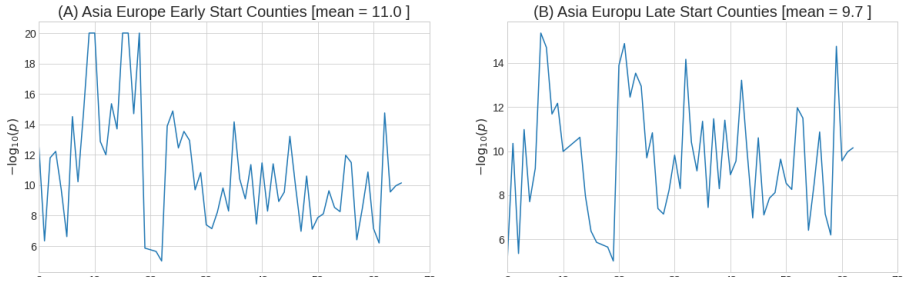

Figure 5. Goodness-of-fit assessment of the linear model of epidemiological data for Asia-Europe countries. Vertical axis is the $-\log_{10}(p)$, where p is the p-value for the slope given by linear regression of log-transformed data (that is the exponential growth of the beginning curve of infected people). The countries are on the horizontal axis in alphabetical order.

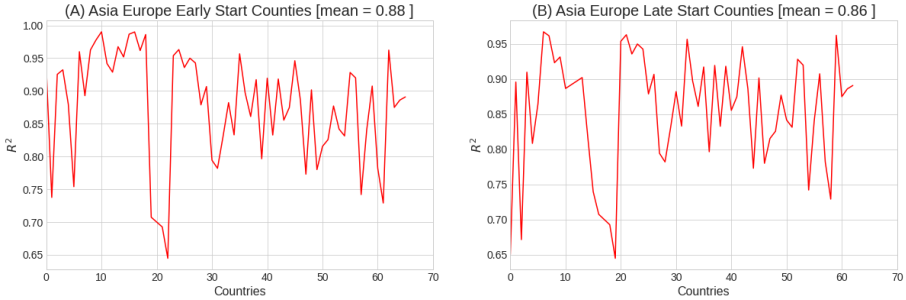

Figure 6. Goodness-of-fit assessment of the linear model of epidemiological data for Asia-Europe countries. Vertical axis is the R^2 score of the linear regression of log-transformed data (that is the exponential growth of the beginning curve of infected people). The countries are on the horizontal axis in alphabetical order.

DISCUSSION

The first confirmed COVID-19 case in the United States was in the state of Washington on January 20, 2020 (Centers for Disease Control and Prevention, 2020; Holshue et al., 2020). This would explain the similarities in transmission processes in the US West coast when compared to China, Japan and Taiwan. On the other hand, on the US East coast, especially New York, the likelihood of the beginning of the COVID-19 epidemic is of European origin.

We analyzed the frequencies of D614 and G614 variants in the US West and East and Asia-Europe West and East. We have shown that irrespective of initial frequencies of these variants at early epidemic stages, G614 always predominates, and very quickly either becomes fixed or significantly surpasses D614 after 60 days of the initial infection. In a previous study (Korber et al., 2020) the conjecture of founder effects and selection of D614 and G614 are compared. Our analysis (Figure 1) indicates that at early epidemic stages the founder effect is more prominent and at later stages selection has an increased impact.

The population dynamics analysis (Figure 2), indicates a selective sweep of G614 over D614. A selective sweep is a population genetics process in which a novel beneficial mutation increases its frequency to a point where it reaches 100% and is therefore “fixed” in

the population (Hermisson and Pennings, 2005). The current COVID-19 epidemics and the discovery of a spike protein variant with a mutation from D to G at position 614 has given an opportunity to show such process in action due to the detailed epidemiological data and viral genome sequence availability. Our analysis also shows that the founder effect and selective sweep are not specific to a country or region, which suggests that the selective advantage of G614 over D614 is global and occurs irrespective of the genetic variation and ethnic background of the host populations.

Although the results indicate that there is a robust difference between the D614G variant and the epidemic growth rate curve it is important to point there are several mutations occurring in the viral genome, which could result in a mutation balance in the viral fitness. That is, it is unlikely that the fitness increase by G614 alone drove the epidemic curves. As shown by (Korber et al., 2020) other mutations hitchhike around G614 by recombination and therefore the combined fitness of several mutations increase the fitness of SARS-CoV-2 to a point that explains the increase observed (Figure 2). Although one study proposes that there is no evidence for increased transmissibility from recurrent mutations in SARS-CoV-2 (Dorp et al., 2020), others suggest a localized transmission of minority variants (Lythgoe et al., 2020) and the increased infectivity of D614G mutation in the SARS-CoV-2 spike protein due to reduced S1 shedding (Zhang et al., 2020). Nevertheless, we provide population study evidence for the hypothesis that combines a founder effect with selection. We hypothesize that G614 predominance over D614 is an example of selective sweep in a viral population (Kang et al., 2021).

With our method, the initial segment of epidemiological curves (of new cases), at early and late epidemic stages, are compared. At early stages, when containment measures were still not fully implemented, the variants are supposed to be unconstrained, and their growth curves might approximate the free viral dynamics characterized by an exponential growth. At later epidemic stages, the implementation of containment measures does not affect our analysis since, currently, the predominant variant has already ‘taken over’ the genetic landscape. The method we used contrasts with other methods to estimate transmission rates and selective advantage because it does not rely on extensive genomic sequencing (Dorp et al., 2020; Lythgoe et al., 2020; Trucchi et al., 2021; Zhang et al., 2020)

We applied our method to the temporal distributions of the SARS-CoV-2 samples bearing D or G at position 614, in two geographic scenarios: USA (East Coast versus West Coast) and Europe-Asia (East Countries versus West Countries). The analysis reveals that D614G prevalence and the growth rates of COVID-19 epidemic curves are correlated at the early stages and not correlated at the later stages, in both the USA and Europe-Asia scenarios. These results suggest that a selective advantage of D614G, given by a higher transmissibility of the new variant, can explain, at least in part, the increased propagation of this variant. This method reveals the effect of a new variant, with higher transmissibility, which can be detected in the host population at countrywide and worldwide levels. Finally, we emphasize that our method does not depend on extensive whole-genome sequencing. The method here presented, requires only the relative frequencies of the variants, which can be obtained by less expensive methods, such as, the most common variant detection data by RT-qPCR (Vogels et al., 2021; Wang et al., 2021). Therefore, an important feature of our method is that it can be applied using much less expensive surveillance technologies, especially in locations with limited genomic sequencing capability.

AUTHOR CONTRIBUTIONS

IMVGC, LMRJ, MRSB, FA: Data analysis planning and conceptualization. TNF, FA: Performing the data analysis. All authors: writing and editing the manuscript.

ACKNOWLEDGMENTS

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, grant 2020/08943-5 to L.M.J, M.R.S.B and F.A. and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, grant 303912/2017-0 to M.R.S.B.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Arons MM, Hatfield KM, Reddy SC, Kimball A, et al. (2020). Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. *N. Engl. J. Med.* 382: 2081-2090. doi: 10.1056/NEJMoa2008457

Centers for Disease Control and Prevention (2017). CDC - Severe Acute Respiratory Syndrome (SARS). https://www.cdc.gov/sars/. Accessed 11 Jun 2021.

Centers for Disease Control and Prevention (2019). CDC - Middle East Respiratory Syndrome (MERS). cdc.gov/coronavirus/mers/about/index.html. Accessed 11 Jun 2021.

Centers for Disease Control and Prevention (2020). First Travel-related Case of 2019 Novel Coronavirus Detected in United States. https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html. Accessed 11 Jun 2021.

Centers for Disease Control and Prevention - Department of Health and Human Services (2004). CDC - Severe Acute Respiratory Syndrome - Fact Sheet: Basic Information about SARS. 3.

Dorp L van, Richard D, Tan CC, Shaw LP et al. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *bioRxiv* 2020.05.21.108506. doi: 10.1101/2020.05.21.108506.

Drosten C, Günther S, Preiser W, van der Werf S, et al. (2003). Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N. Engl. J. Med*. 348: 1967-1976. doi: 10.1056/NEJMoa030747.

Elbe S and Buckland-Merrett G (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. *Global Challenges*. 1: 33-46. doi: 10.1002/gch2.1018.

Fam BSO, Vargas-Pinilla P, Amorim CEG, Sortica VA, et al. (2020). ACE2 diversity in placental mammals reveals the evolutionary strategy of SARS-CoV-2. *Genet. Mol. Biol.* 43:e20200104. doi: 10.1590/1678-4685-gmb-2020-0104.

Gandhi M, Yokoe DS and Havlir DV (2020). Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *N. Engl. J. Med.* 382: 2158-2160. doi: 10.1056/NEJMe2009758.

Hermisson J and Pennings PS (2005). Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics*. 169: 2335-2352. doi: 10.1534/genetics.104.036947.

Hijawi B, Abdallat M, Sayaydeh A, Alqasrawi S, et al. (2013). Novel coronavirus infections in Jordan, April 2012: epidemiological findings from a retrospective investigation. *East Mediterr. Health J.* 19 Suppl. 1: S12-18.

Holshue ML, DeBolt C, Lindquist S, Lofy KH et al. (2020). First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* 382: 929-936. doi: 10.1056/NEJMoa2001191.

Johns Hopkins University (2020). COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. Accessed 10 Jun 2021.

Kang L, He G, Sharp AK, Wang X, et al. (2021). A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. *Cell*. 184: 4392-4400.e4. doi: 10.1016/j.cell.2021.07.007.

Korber B, Fischer W, Gnanakaran S, Yoon H, et al. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *Cell.* 182: 812-827. doi: 10.1101/2020.04.29.069054.

Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, et al. (2003). A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* 348:1953–1966. doi: 10.1056/NEJMoa030781.

Lee N, Hui D, Wu A, Chan P, et al. (2003). A Major Outbreak of Severe Acute Respiratory Syndrome in Hong Kong. *N. Engl. J. Med.* 348: 1986-1994. doi: 10.1056/NEJMoa030685.

Los Alamos National Laboratory (2020). SARS-CoV-2 Sequence Analysis pipeline - SARS-CoV-2 map: Distribution of D614 and G614. https://cov.lanl.gov/apps/covid-19/map/. Accessed 11 Jun 2020.

Lythgoe KA, Hall M, Ferretti L, Cesare M de, et al. (2020). Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. *bioRxiv*. 2020.05.28.118992. doi: 10.1101/2020.05.28.118992.

Peiris J, Lai S, Poon L, Guan Y, et al. (2003) Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet*. 361: 1319-1325. doi: 10.1016/S0140-6736(03)13077-2.

Shu Y and McCauley J (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*. 22: 30494. doi: 10.2807/1560-7917.ES.2017.22.13.30494.

Spiegelhalter DJ (1986). Probabilistic prediction in patient management and clinical trials. *Stat. in Med.* 5: 421-433. doi: 10.1002/sim.4780050506.

Trucchi E, Gratton P, Mafessoni F, Motta S, et al. (2021). Population Dynamics and Structural Effects at Short and Long Range Support the Hypothesis of the Selective Advantage of the G614 SARS-CoV-2 Spike Variant. *Mol. Biol. Evol.* 38: 1966-1979. doi: 10.1093/molbev/msaa337.

United States Census Bureau (2020b). https://www.census.gov/programs-surveys/acs. Accessed 11 Jun 2021.

U.S. Bureau of Labor Statistics (2020c). https://www.bls.gov/. Accessed 11 Jun 2021.

Vogels CBF, Breban MI, Ott IM, Alpert T, et al. (2021). Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLOS Biology* 19:e3001236. doi: 10.1371/journal.pbio.3001236.

Wang H, Miller JA, Verghese M, Sibai M, et al. (2021). Multiplex SARS-CoV-2 Genotyping PCR for Population-Level Variant Screening and Epidemiologic Surveillance. *medRxiv*. 2021.04.20.21255480. doi: 10.1101/2021.04.20.21255480.

World Health Organization (2020a). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. In: World Health Organization. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 10 Jun 2021.

World Health Organization (2020b). WHO Timeline - COVID-19. In: World Health Organization. https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19. Accessed 10 Jun 2021.

World Health Organization (2020c). Coronavirus disease 2019 (COVID-19) *Situation Report* - 94. 12.

World Health Organization (2004). China's latest SARS outbreak has been contained, but biosafety concerns remain – Update 7. In: World Health Organization. https://www.who.int/csr/don/2004_05_18a/en/. Accessed 11 Jun 2021.

World Health Organization (2019). Middle East respiratory syndrome coronavirus (MERS-CoV) - Key Facts. https://www.who.int/en/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov). Accessed 11 Jun 2021.

World Health Organization (2020d). MERS *Situation Update* - January 2020. 1.

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, et al. (2012) Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N. Engl. J. Med.* 367: 1814-1820. doi: 10.1056/NEJMoa1211721.

Zhang L, Jackson CB, Mou H, Ojha A et al. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*. 2020.06.12.148726. doi: 10.1101/2020.06.12.148726.

Zhong N, Zheng B, Li Y, Poon L, et al. (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet*. 362: 1353-1358. doi: 10.1016/S0140-6736(03)14630-2.