Analysis of SARS-CoV-2 mutations reveals three types of temporal dynamics and one is correlated with international travels

Santiago Justo Arevalo*[1,2], Daniela Zapata Sifuentes[1], César J. Huallpa[3], Gianfranco Landa Bianchi[1], Adriana Castillo Chávez[1], Romina Garavito-Salini Casas[1], Carmen Sofia Uribe Calampa[1], Guillermo Uceda-Campos[4], Roberto Pineda Chavarría[1]

[1]Facultad de Ciencias Biológicas, Universidad Ricardo Palma, Lima, Peru

[2]Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, Brazil

[3]Facultad de Ciencias, Universidad Nacional Agraria La Molina, Lima, Peru

[4]Facultad de Ciencias Biológicas, Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Peru

*Corresponding author: santiago.justo@urp.edu.pe, santiago.jus.are@usp.br

Abstract

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This disease has spread globally, causing more than 161.5 million cases and 3.3 million deaths. Keeping on the identification, surveillance and the study of the temporal dynamics of mutations with significant representation is central to understand the adaptation of SARS-CoV-2. Furthermore, how lockdown policies influence the dynamics of SARS-CoV-2 mutations is poorly understood. Here, using 1 058 020 SARS-CoV-2 genomes and COVID-19 cases from 714 country-month combinations representing 98 countries, we performed a normalization by COVID-19 cases calculation of relative frequency of SARS-CoV-2 mutations. We found 115 mutations estimated to be present in more than 3 % of global COVID-19 cases and determined three types of mutation dynamics: High-Frequency, Medium-Frequency, and Low-Frequency. Classification of mutations based on temporal dynamics helps to study viral adaptation and can be used to evaluate the effects of human behaviors in the pandemic. For instance, we report a negative and positive correlation of the frequency change of High-Frequency mutations with the level of international movement controls and the number of flight departures, respectively.

Introduction

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a single-stranded positive RNA virus that infects humans. Since the first reported cases in December 2019, the disease has spread globally causing more than 161.5 million confirmed cases and 3.3 million deaths as of May 16th [1].

Since the emergence of COVID-19, massive genome sequencing has played a central role in the analysis and understanding of the evolutionary dynamics of the virus. This has allowed the identification of mutations that appeared early in the pandemic and now seem to be fixed in the population [2,3,4,5,6]. In addition, the effectiveness of vaccines against variants coding for mutations in the spike protein is being carefully monitored [7,8,9,10,11,12]. Therefore, the identification and surveillance of mutations with significant representation in the population is of utmost importance in the fight against this pandemic.

The combination of virus spread by droplets through close contact, the large number of asymptomatic cases, the absence of effective pharmaceutical treatments at the beginning of the pandemic and the delays in production and distribution vaccines, leave non-pharmaceutical interventions as the most effective measures to contain the spread of COVID-19 for a large fraction of the world´s population.

Different studies have evaluated the relationship between non-pharmaceutical interventions and the decrease in the number of cases [13,14], the reproductive number [13,15,16], the case fatality rate [17], the contagion rate [18], and the number of SARS-CoV-2 importations [19,20]. On the other hand, the effect of non-pharmaceutical interventions, particularly those related to an international movement, on mutation frequency variation has been less analyzed. Pachetti et al. (2020) analyzed how lockdown policies might have influenced the dynamics of some SARS-CoV-2 mutations; however, no quantitative or semiquantitative indicator is shown that could define more precisely the reported effect. Muller et al. (2021) use phylogenetic methods to estimate the importance of SARS-CoV-2 introductions on increasing the relative frequency of the D614G mutation, implicitly showing that international movement can affect the relative frequency of mutations.

Here, we analyzed the temporal dynamics of SARS-CoV-2 mutations from 1 058 020 genomes estimated to be present in more than 3 % of global COVID-19 cases. We then investigated whether there is a correlation between international travel and the increase in the frequency of the most prevalent mutations.

Results:

115 mutations overpass presence on 3 % of global COVID-19 cases and most of them are non-synonymous:

We performed a by case normalization of the frequencies of the mutations from 1 058 020 genomes all around the world. By case normalization represents the frequency of cases where a mutation is present. We named this normalized relative frequency as NRFp. The NRFp of each mutation was calculated from genomes and the number of cases of 714 country-month combinations, including 98 countries from January 2020 to April 2021.

We identified 115 mutations with NRFp > 0.03 (Fig. S1); this means that those mutations are estimated to be present in more than 3 % of the COVID-19 cases globally. Considering that the sum of the reported cases from the 714 country-month combinations analyzed was 120 008 410 cases, an NRFp of more than 0.03 means that those mutations were present in more than 3 600 252 global COVID-19 cases.

From these 115 mutations, one is present in an intergenic region, two are present in the 5'-UTR, 22 are synonymous and 90 are non-synonymous (Table 1). This results in a dN/dS ratio of 4.1 that could imply positive selection occurring in the SARS-CoV-2 genome. Intriguingly, the 22 synonymous mutations are all found in non-structural proteins, and just 25 of the 90 non-synonymous mutations fall in non-structural proteins.

Furthermore, some of the remaining 65 non-synonymous mutations on the structural proteins are widely reported, such as Δ69-70, A222V, E484K, N501Y, and D614G on spike (S) protein, and R203K, G204R on nucleocapsid (N) protein. 48 out of 65 non-synonymous mutations lie on the S and N protein and just 17 on other structural proteins (7 in orf8, 4 in orf3a, 3 in membrane (M) protein, 2 in orf6, and 1 in orf10) (Table 1). Therefore, ~74 % of the non-synonymous mutations are found in S or N proteins suggestive of positive selection.

Mutations show three types of temporal dynamics:

The dynamics of the 115 mutations were analyzed calculating the NRFp in each month from January 2020 to April 2021 (Fig. 1). We observed three types of dynamics: 1) Mutations that present frequency changes between two months greater than 0.15 and did not present frequency changes between two months lesser than -0.05 were named "high-frequency" (HF) mutations (Fig. 1A-C). 2) Not HF mutations estimated to be present in 10 % of the cases (0.1 NRFp) in at least one month were classified as "medium-frequency" (MF) mutations (Fig. 1D-F). 3) Mutations without months representing more than 10 % of the global COVID-19 cases were named as "low-frequency" (LF) mutations (Fig. 1G-I).

HF mutations are characterized by a rapid increase in global frequency since its appearance (Fig. 1A-C) that could be determined by positive selection without competition and/or by other effects related to population dynamics such as absence of international movement controls.

Those mutations appeared in two well-defined stages of the pandemic. The first group is composed of four mutations that now are apparently globally fixed. They emerged at the beginning of the pandemic in January 2020, reaching more than 0.75 NRFp in April 2020 (Fig.

1A-C, January-2020 to June-2020). The second group surged in December 2020 and now (April 2021) have surpassed the 0.5 NRFp (Fig. 1A-C, November-2020 to April-2021).

The first and second groups contained Spike mutations well known due to their possible implications in vaccine efficacy such as D614G, on the first group, and Δ69-70, N501Y, on the second group (Fig. 1B). However, the individual effects of other mutations are not well reported in the literature such as Δ144, A570D, P681H, T716I, S982A, D1118H on S protein, mutations on non-structural proteins (nsp3, nsp6, nsp12) (Fig. 1A), and mutations on the ORF8 and the N protein (Fig. 1B). It is important to note that all the mutations belonging to the second group are those that define B.1.1.7 lineage.

Most of the MF mutations showed a pattern of several months with a roughly constant NRFp, some of them with months showing an increase or decrease in the relative frequency (Fig. 1D-F). This could mean beneficial mutations but with a competition that avoids their rapid fixation on the population. Nucleocapsid mutations G28881A, G28882A, and G28883C clearly show this pattern, with a steady frequency increase up to July 2020, followed by a decrease in global frequency and a new increase since December 2020 (Fig. 1E).

Several non-synonymous mutations in non-structural proteins (nsp2, nsp6, nsp12, and nsp13) (Fig. 1D) and structural proteins (S, ORF3a, M, ORF8, N, ORF10) (Fig. 1E) corresponds to this type of dynamic, including the well-reported E484K in S protein.

A pattern similar to MF mutations is seen in LF mutations but with frequencies below 0.1 in all the months for both non-synonymous mutations in non-structural proteins (Fig. 1G) and structural proteins (Fig. 1H). Most of those mutations have detectable frequencies in several months that could imply that, despite their low frequency, they have biological implications in SARS-CoV-2.

Three types of mutations are described here and named “High-Frequency” (HF), “Medium-Frequency” (MF), and “Low-Frequency” (LF) according to their temporal dynamics. The three types are represented in all the months analyzed, and found in non-structural and structural proteins. Differences in the dynamics could be due to different forces acting on each of these types of mutations and the implications of all of them need to be studied to better understand the adaptation process of SARS-CoV-2.

Flight departures and international movement controls are partially correlated with a frequency change in High-Frequency mutations

The rapid increase in global frequency of HF mutations and the observation that those mutations appeared at two very defined stages of the pandemic (Fig. 1A-C) lead us to hypothesize that, at least part of this abrupt increase is due to changes in the international migration patterns.

To test this hypothesis, we used two indicators: 1) Level of international travel controls obtained from Oxford COVID-19 Government Response Tracker (OxCGRT), and 2) Number of flight departures from the International Civil Aviation Organization (ICAO) to correlate the change in frequency of HF mutations and the level of international travel controls or the number of flight departures (see material and methods for a detailed description of how data collection).

To better analyze the correlations, we grouped information of the two indicators into six regions: Africa, Asia, Europe, North America, Oceania, and South America. As shown in Figure 2A, policies of international travel controls were first implemented during the first third of 2020, having a peak (the highest level) in April 2020 in all six regions (Fig. 2A). This was followed by changes in the level of controls depending on the region analyzed (Fig. 2A).

Some countries still allowed international flights while policies of international travel controls were active. Therefore, the number of flight departures can be used as a complementary indicator of international movement within a region (Fig. 2B). The pattern between policies of international travel controls and the number of departures is positively correlated (Fig. S2); and so these two indicators were used to measure the correlation of international movement with the change of frequency of HF mutations.

The mean change of frequency of the HF mutations was calculated by month in each of these regions (Fig. 2A-B), and the correlation with the international travel controls (Fig. 2C) or with the number of flight departures (Fig. 2D) was measured.

A moderate but significant ($p < 0.05$) negative correlation was observed between the level of international travel controls and frequency change of HF mutations. This means that months where policies of international travel controls were active, the increase of frequency of HF mutations was less than when those policies were not active. Moreover, the relative number of departures showed a significant positive correlation with frequency change; also showing that more flight departures correspond with a greater increase in the frequency of HF mutations.

Discussion

Using a large dataset of 1 058 020 genomes of 714 country-month combinations that include 98 countries from six geographical regions, and the number of cases reported from these country-month combinations, we were able to estimate the number of cases where a SARS-CoV-2 mutation was present. The normalized relative frequency analysis of mutations allowed us to identify 115 mutations, each of which is found in at least 3.6 million COVID-19 cases, most of them belonging to the S and N proteins (Table 1).

The relative frequency normalization by cases allowed us to identify mutations that have not been reported in other global studies, such as that of Castonguay et al. 2021. This is because in many countries the number of genomes sequenced is low and certain mutations could go unnoticed. This highlights the importance of normalization in global analyses.

Three types of temporal dynamics are observed for those mutations. High-Frequency mutations (HF) that rapidly reach more than 50 % global relative frequency, Medium-Frequency mutations (MF) that maintains a roughly constant global relative frequency of more than 10 % during several months, and Low-Frequency mutations (LF) that also maintain their frequencies between but below 10 %.

HF mutations identified here have been widely reported [23] due to their presence in variants of concern. We describe the temporal dynamics of these mutations showing that they are characterized by a rapid increase of frequency over periods of few months (Fig. 1A-C). This finding highlights the importance of rapid identification of other mutations of concern and characterization on their dynamics.

Some of the MF and LF mutations that we observed have been less reported, described only in particular countries or regions [24,25,26], or not reported, such as K997Q on nsp3 and S202C on N protein. We think that it is also important to study MF and LF mutations because they are present in several months throughout the pandemic and we did not observe the extinction of any of these mutations (relative frequency of 0 or near to 0 in two or more months) (Fig. 1D-I). Some benefits may be conferred to SARS-CoV-2 by those mutations that prevent their extinction, but constant competition avoids its rapid increase in frequency [27,28].

Climatological factors influence propagation and mortality rates of SARS-CoV-2, and also the frequency of specific mutations [29,30,31,32]. A global, roughly constant, relative frequency of MF and LF mutations could be explained by their persistence in regions with specific

climatological factors. In our previous work [33], we observed that the mutation T85I in nsp2 has much more frequency in North America than in other continents. Here, we showed that this MF mutation maintains this tendency persisting since its appearance at ~10 % NRFp (Fig. 1D).

Mutation E484K on S protein is present in different variants, such as B.1.351 (South Africa) and B.1.1.28 (Brazil), which has been interesting for being in the receptor-binding domain of spike protein and its relationship with the evasion of the immune response [9]. Intriguingly, despite having appeared in the same month of the second group of the HF mutations (those that characterize lineage B.1.1.7), this mutation corresponds to the MF type of dynamic (Fig. 1E). This could be due to the limited effect that it has when it is not in combination with other spike mutations [34].

The study of the mutation dynamics permits us to study the patterns of adaptation of viruses on a global scale and evaluate the effect of human population behaviors on this process. Gathering the information of temporal dynamics of HF mutations together with publicly available information of human behavior (in this case international travel patterns), we found a correlation between the frequency variations of several mutations with the level of the international movement.

Correlation analyses between the mean of HF mutations frequency changes and the level of international travel controls or the number of flight departures have revealed a negative and positive correlation between them, respectively. The moderate level of correlation may be because international movement is only one of several factors for the rapid increase in the frequency of those mutations. A combination of positive selection for those mutations and the rapid spread favored by international travel could permit the currently high frequencies of those mutations.

Non-pharmaceutical interventions have been shown to reduce the number of COVID-19 cases [13], the contagion rate [18], the reproductive number [13,15,16], the case fatality rate [17] and the number of SARS-CoV-2 introductions and/or exportations [19,20,35]. Here we showed that international movement is negatively correlated with frequency change of HF mutations (Fig. 2C). To our knowledge, this is the first report that relates a non-pharmaceutical intervention with the increase in the frequency of specific mutations in a global context.

A future challenge is to investigate the implications of the three observed dynamics of mutation and inquire how MF and LF mutations participate in the adaptation process of SARS-CoV-2. It would also be of considerable interest to develop new approaches on how we can determine the effects of human behavior on the adaptation dynamics of SARS-CoV-2.

Limitations of the study

International travel controls obtained from OxCGRT are based on a semiquantitative criterion that embraces just 5 levels. Due to the variations in sociocultural and economic factors among different countries, this criterion does not have enough resolution to reveal small differences between countries on international movement controls policies. However, it allows us to have a general global vision at a regional level.

Unfortunately, to our knowledge, there are few publicly available databases where we can obtain detailed data about international movement. The number of flight departures may not be the best indicator of how many people traveled, but it gave us an approximate.

Although the methodology of normalization by cases alleviates the differences in the number of genomes sequenced by country, confidence in the calculation of relative frequencies of mutations is still low in regions with a low number of genomes sequenced. For example, a mutation with 0.5 relative frequency that comes from a sample of 15 genomes will have a

confidence interval between 0.25 and 0.75; on the other hand, a sample of 150 genomes will generate a confidence interval between 0.58 and 0.42. Moreover, the number of cases is still subjected to bias due to for instance, the difference in the number of tests that each country performs.

Conclusions

Normalization by cases of the frequency of mutations is an important tool for global analyses in a pandemic where not all the countries possess the same capacity to sequence SARS-CoV-2 genomes. This process alleviates these differences but does not eliminate this problem. Worldwide efforts to help countries with fewer sequencing resources would improve our understanding of the adaptation and evolution process of SARS-CoV-2

Three types of dynamics of mutations were identified here, High-Frequency, Medium-Frequency, and Low-Frequency mutations. Description of these types of dynamics helps in the understanding of the adaptation process of SARS-CoV-2.

Different from controlled experiments on evolution, the massive global sequencing of SARS-CoV-2 genomes allows us to investigate the adaptation process of a virus in the real world. In this pandemic, human behavior can influence the adaptative process of SARS-CoV-2. Correlation of the increment of frequency of High-Frequency mutations with international travel could be evidence for this.

Methods

Normalized by cases relative frequency of mutations on the SARS-CoV-2 genome:

To perform mutation frequency analysis considering the number of cases in each country we followed similar steps as described in Justo et al. 2021 [33], with some modifications: we first downloaded 1 221 746 genomes from the GISAID database (as of April 24th, 2021). Sequences with less than 29 000 nt were removed and the resulting sequences were aligned against the reference SARS-CoV-2 genome (EPI_ISL_402125) from nt 203 to nt 29674 using ViralMSA.py [36,37]. From this alignment, we removed sequences with more than 290 Ns, more than 0.05 % unique mutations, and/or more than 2 % gaps. After those filters, we had 1 058 020 genomes. Subalignments were generated by grouping sequences by country and month. Subalignments with less than 15 sequences were not considered in the analysis. Nucleotide relative frequencies of each genomic position on each of 714 subalignments each corresponding to a different country-month combination (including 98 countries) were calculated. Relative frequencies were multiplied by the number of cases reported on the specific country-month. The number of cases for each month and country was obtained from the European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide). To obtain the global estimates of the relative frequency of mutations in COVID-19 cases, we added the previous multiplications and divide them by the total number of cases of the 714 country-month combinations. These relative frequencies mean the percentage of global COVID-19 cases where a particular mutation is present was named Normalized Relative Frequency by position (NRFp). The same procedure was done to obtain the NRFp of the mutations by months. Data manipulation was done using R and python scripts.

Calculations of Frequency Change of HF mutations:

The frequencies by country-months of each HF mutation were obtained from the previous calculation. Then, we calculated the mean of the normalized relative frequencies by region-month (Africa, Asia, Europe, North America, Oceania, South America) of each HF mutation. Each mean of the normalized relative frequency of each region-month was subtracted to the

normalized relative frequency of the previous month to obtain the Frequency change of each HF mutation by month by region. Thus, the frequency change of February 2020 corresponds to the subtraction of the normalized relative frequency of March 2020 minus the normalized relative frequency of February 2020. These subtractions follow the logic that if a policy is implemented at the beginning of March, the effects on the frequency change will probably be seen during this month until the next month. Finally, frequency changes lesser than 0 were removed and the mean of the frequency change of the mutations was calculated. Standard deviations of those frequency changes were reported as the frequency change variations of different HF mutations within a region.

Calculation of Relative Level of International Travel controls:

Information on international travel controls was obtained from Oxford COVID-19 Government Response Tracker (OxCGRT) (https://ourworldindata.org/covid-international-domestic-travel). This parameter includes record restrictions by country on international travel on an ordinal scale: 0 – No measures, 1 – Screening, 2 – Quarantine arrivals from high-risk regions, 3 – Ban on arrivals from some regions, 4 – Ban on all regions or total border closure. Using the daily information of each of 697 country-month combinations of the 714 used on the NRFp calculation available in this webpage, we calculated a mean of this ordinal scale. After that, we convert this mean to a relative frequency taking 4 as 100 %. Finally, we calculated the mean of this relative level of international level controls by region (Africa, Asia, Europe, North America, Oceania, South America) and month. Standard deviations were reported as the intra-region variation.

Calculation of Relative Number of Daily Flight Departures:

Information on the number of flight departures was obtained from the International Civil Aviation Organization (ICAO) (https://www.icao.int/safety/pages/covid-19-airport-status.aspx). The average of daily departures by week was obtained from available country-month combinations in this webpage (672 of the 714 used on the NRFp calculation). The mean of daily departures of each country-month was calculated. Information was grouped by region (Africa, Asia, Europe, North America, Oceania, South America) and month calculating the mean for each region-month. Standard deviations were reported as the intra-region variation.

Correlations of the relative frequency change of HF mutations with the level of international travel controls and the relative number of daily flight departures:

Analysis of Spearman correlations by region-month of the relative frequency change of HF mutations and the relative level of international travel controls or the relative number of daily flight departures were performed using R [38] and the packages ggplot2 [39] and ggpubr [40]. P values lesser than 0.05 were considered significant.

References

1. World Health Organization (2021). Available at: https://covid19.who.int/ (Accessed May 16th, 2021).
2. Korber, B. et al. Tracking Changes in SARS-Cov-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182, 812-827 (2020).
3. Biswas, S. K., & Mudi, S. R. Spike protein d614g and RDRP p323l: The SARS-CoV-2 mutations associated with severity of COVID-19. Genomics Inform. 18(4), 1–7; 10.5808/GI.2020.18.4.e44 (2020).
4. Korber, B., et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. Preprint at https://doi.org/10.1101/2020.04.29.069054 (2020).

5. Yurkovetskiy, L. et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Cell, 183(3), 739-751; 10.1016/j.cell.2020.09.032 (2020).
6. Callaway, E. Making sense of coronavirus mutations. Nature 585, 174-177 (2020).
7. Khan, A., et al. Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data. J. Cell. Physiol. 1–13; 10.1002/jcp.30367 (2021).
8. Tegally, H. et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. Preprint at https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1 (2020).
9. Jangra, S. et al. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. Lancet Microbe; 10.1016/S2666-5247(21)00068-9 (2021).
10. Leung, K., Shum, M. H. H., Leung, G. M., Lam, T. T. Y., & Wu, J. T. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Eurosurveillance 26(1); 10.2807/1560-7917.ES.2020.26.1.2002106 (2021).
11. Liu, Y. et al. The N501Y spike substitution enhances SARS-CoV-2 transmission. Preprint at https://doi.org/10.1101/2021.03.08.434499 (2021).
12. Kemp, S. et al. Recurrent emergence and transmission of a SARS-CoV-2 spike deletion H69/70. Preprint at https://www.biorxiv.org/content/10.1101/2020.12.14.422555v6 (2021).
13. Hyafil, A. & Moriña, D. Analysis of the impact of lockdown on the reproduction number of the SARS-CoV-2 in Spain. Gac Sanit; 10.1016/j.gaceta.2020.05.003 (2020).
14. Hsiang, S. et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. Nature, 584(7820), 262-267 (2020).
15. Agrawal, M., Kanitkar, M., & Vidyasagar, M. Modelling the spread of SARS-CoV-2 pandemic-Impact of lockdowns & interventions. Indian J Med. Res. 153, 175-181 (2021).
16. Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., & Jit, M. The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. BMC Med. 19(1), 1-12 (2021).
17. Pachetti, M. et al. Impact of lockdown on Covid-19 case fatality rate and viral mutations spread in 7 countries in Europe and North America. J. Transl. Med. 18(338) (2020).
18. Larrosa, J. M. SARS□CoV□2 in Argentina: Lockdown, mobility, and contagion. J Med Virol. 93(4), 2252-2261 (2021).
19. Adekunle, A., Meehan, M., Rojas-Alvarez, D., Trauer, J. & McBryde, E. Delaying the COVID-19 epidemic in Australia: evaluation of the effectiveness of international travel bans. Aust NZ J Public Health 44(4), 257-259; 10.1111/1753-6405.13016 (2020).
20. Wells, C. R. et al. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. Proc Natl Acad Sci. 117(13), 7504-7509 (2020).
21. Muller, N. F. et al. Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. Sci. Transl. Med 10.1126/scitranslmed.abf0202 (2021).

22. Castonguay, N., Zhang, W. & Langlois, M. Meta-Analysis and Structural Dynamics of the Emergence of Genetic Variants of SARS-CoV-2. Preprint at https://www.medrxiv.org/content/10.1101/2021.03.06.21252994v2 (2021).
23. Plante, J. et al. The variant gambit: COVID-19's next move. Cell Host & Microbe 29 (2021).
24. Hodcroft, E. et al. Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677. Preprint at https://www.medrxiv.org/content/10.1101/2021.02.12.21251658v3 (2021).
25. Nagy, Á., Pongor, S. & Györffy, B. Different mutations in SARS-CoV-2 associate with severe and mild outcome. International Journal of Antimicrobial Agents 57, 106272; 10.1016/j.ijantimicag.2020.106272 (2021).
26. Zrelovs, N. et al. First Report on the Latvian SARS-CoV-2 Isolate Genetic Diversity. Front. Med. 8, 626000; 10.3389/fmed.2021.626000 (2021).
27. Good, B., McDonald, M., Barrick, J., Lenski, R. & Desai, M. The dynamics of molecular evolution over 60,000 generations. Nature 551, 45-50 (2017).
28. Luksza, M. & Lässig, M. A predictive fitness model for influenza. Nature 507, 57-61 (2014).
29. Sharif, N. et al. Impact of meteorological parameters and population density on variants of SARS-CoV-2 and outcome of COVID-19 pandemic in Japan. Epidemiol. Infect. 149, 1- 45. (2021).
30. Sharif, N. & Dey, S. K. Impact of population density and weather on COVID-19 pandemic and SARS-CoV-2 mutation frequency in Bangladesh. Epidemiol. Infect. 149. (2021)
31. Sharif, N., et al. Environmental correlation and epidemiologic analysis of COVID-19 pandemic in ten regions in five continents. Heliyon, 7(3). (2021).
32. Bashir, M. F., et al. Environmental quality, climate indicators, and COVID-19 pandemic: insights from top 10 most affected states of the USA. Environ. Sci. Pollut. Res., 1-10. (2021).
33. Justo, S., et al. Global geographic and temporal analysis of SARS-CoV-2 haplotypes normalized by COVID-19 cases during the pandemic. Front. Microbiol. 12; 10.3389/fmicb.2021.612432 (2021).
34. Collier, D. et al. Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. Nature. 593, 136-141 (2021)
35. Yang, J. et al. Uncovering two phases of early intercontinental COVID-19 transmission dynamics. J. Travel Med. 27, 8; 10.1093/jtm/taaa200 (2020).
36. Moshiri, N. ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes. Bioinformatics. 37, 5; 714-716 (2020).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34, 18; 3094-3100 (2018).
38. Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
39. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2016).
40. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. URL https://CRAN.R-project.org/package=ggpubr

Figure Legends:

Table 1. 115 mutations estimated to be present in more than 3 % global COVID-19 cases. ID: code of identification on the format: nucleotide change_aminoacid change_region of mutation. NRFp: Global normalized by COVID-19 cases relative frequency. Type: type of temporal

dynamic of the mutation; High-Frequency (HF), Medium-Frequency (MF), Low-Frequency (LF). nt_change: change at nucleotide level. AA_change: change at aminoacid level. Location: Specific genomic region where the mutation occurred. Syn: Synonymous change

Figure 1. Three different temporal dynamics of SARS-CoV-2 mutations. Normalized by cases Relative Frequency (NRFp) of the 115 identified mutations by month. To improve visualization mutations were divided into three columns. Left, medium and right columns show non-synonymous in non-structural proteins, non-synonymous in structural proteins, and synonymous mutations, respectively. (**a**), (**b**), and (**c**) show mutations with frequency changes higher than -0.05 between all months (High-Frequency mutations). (**d**), (**e**), and (**f**) show not High-Frequency mutations with at least one month with an NRFp greater than 0.1. (**g**), (**h**), and (**i**) show mutations with less than 0.1 NRFp in all months. Error bars represent inter-region variation as standard deviations.

Figure 2. Correlation of relative frequency change of High-Frequency (HF) mutations with international movement indicators. (**a**) Temporal trend of the relative frequency change of HF mutations and relative level of international travel controls in six different regions. (**b**) Temporal trend of the relative frequency change of HF mutations and the relative number of daily flight departures in six different regions. Errors in (**a**) and (**b**) are showed as shades. Error in relative frequency change of mutations means standard deviation between the frequency change of different mutations. Error in the relative level of international travel controls and the relative number of daily flight departures means standard deviation within the region. (**c**) and (**d**) Spearman correlation of relative frequency change of HF mutations with the relative level of international travel controls (**c**) or with the relative number of daily flight departures (**d**). Each point in (**c**) and (**d**) represents a region-month combination. The line represents regression lines and shades show a 95 % confidence interval.

Author contributions

SJA designed the study. SJA, DZS, CHR, GLB, ACC, RG-SC, CSUC, RPC analyzed the data. SJA, CSUC GU-C wrote python and R scripts. The manuscript was written by SJA, DZS, CHR, GLB, ACC, RG-SC, CSUC. All authors discussed the methodologies, results, and read and approved the manuscript.

Acknowledgements

We are very grateful to the GISAID Initiative and all its data contributors, i.e., the authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. To the Ricardo Palma University High-Performance Computational Cluster (URPHPC) managers Gustavo Adolfo Abarca Valdiviezo and Roxana Paola Mier Hermoza at the Ricardo Palma Informatic Department (OFICIC) for their contribution in programs and remote use configuration of URPHPC. Also, to one of the managers of the High-Performance Computer from the Instituto de Investigaciones de la Amazonía Peruana, Rodolfo Cardena Vigo for its assistance in configurations and program installations.

Competing interests:

The authors declare no competing interests.

Data availability:

Publicly available datasets were analyzed in this study. This data can be found at: gisaid.org. Python and R scripts are available upon reasonable request to the corresponding author.

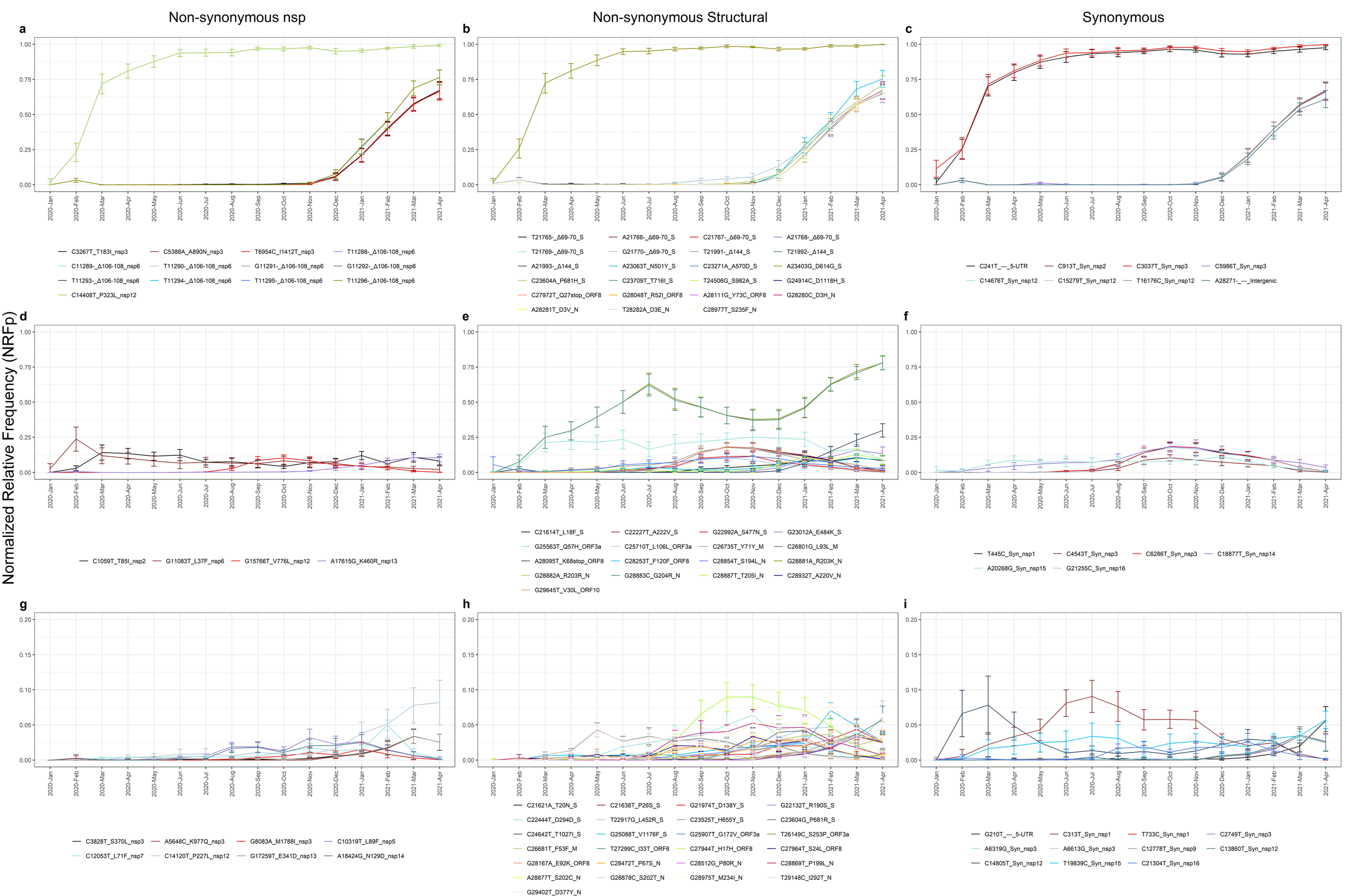
Funding:
We thank to the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) graduate scolarship (to SA; 2015/13318-4) and Universidad Ricardo Palma (URP) for APC financing.

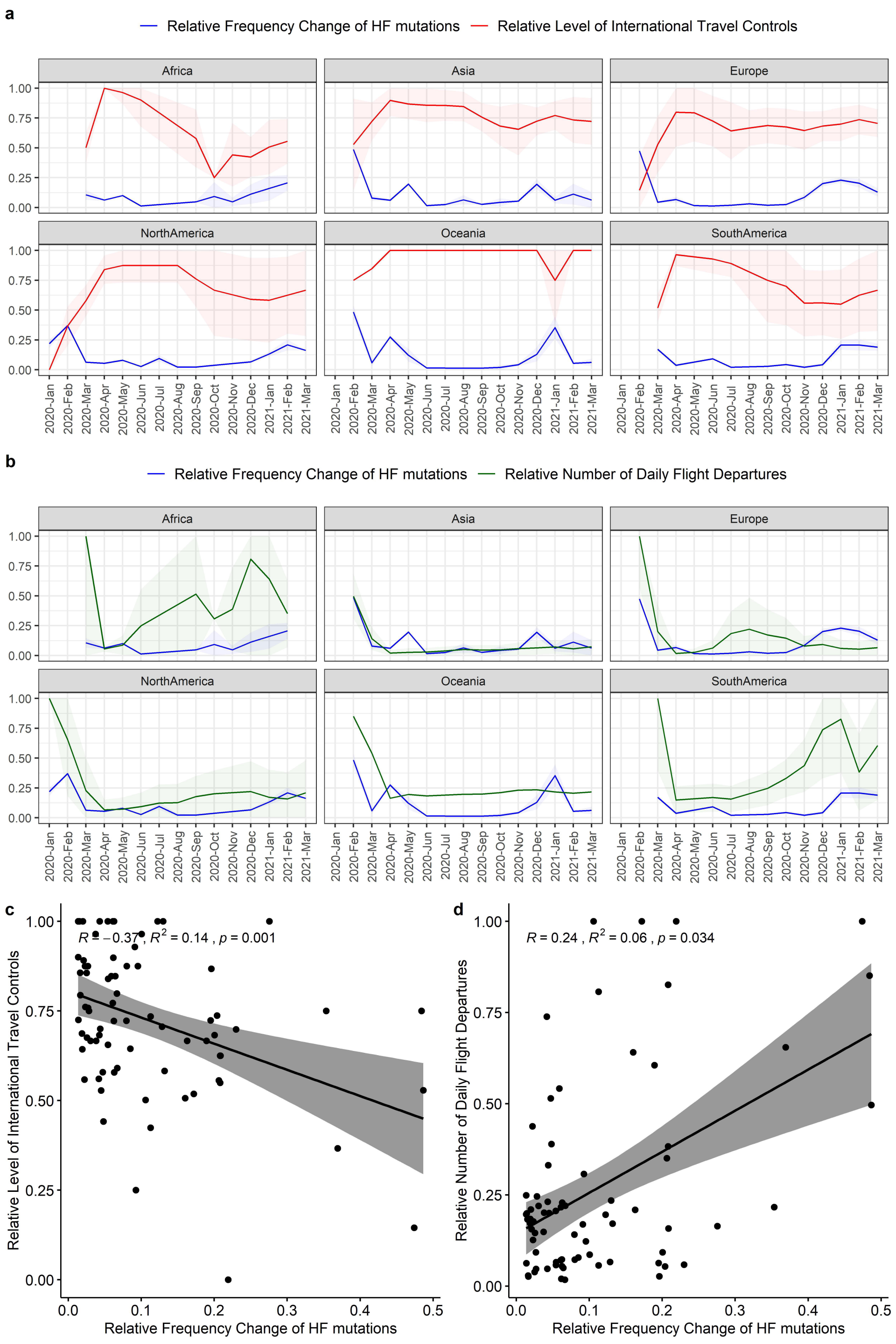
Table 1

ID	NRFp	Type	nt_change	AA_change	Location
G210T_---_5-UTR	0.033	LF	G210T	---	5-UTR
C241T_---_5-UTR	0.971	HF	C241T	---	5-UTR
C313T_Syn_nsp1	0.031	LF	C313T	Syn	nsp1
T445C_Syn_nsp1	0.07	MF	T445C	Syn	nsp1
T733C_Syn_nsp1	0.034	LF	T733C	Syn	nsp1
C913T_Syn_nsp2	0.136	HF	C913T	Syn	nsp2
C1059T_T85I_nsp2	0.194	MF	C1059T	T85I	nsp2
C2749T_Syn_nsp3	0.034	LF	C2749T	Syn	nsp3
C3037T_Syn_nsp3	0.978	HF	C3037T	Syn	nsp3
C3267T_T183I_nsp3	0.153	HF	C3267T	T183I	nsp3
C3828T_S370L_nsp3	0.035	LF	C3828T	S370L	nsp3
C4543T_Syn_nsp3	0.032	MF	C4543T	Syn	nsp3
C5388A_A890N_nsp3	0.136	HF	C5388A	A890N	nsp3
A5648C_K977Q_nsp3	0.034	LF	A5648C	K977Q	nsp3
C5986T_Syn_nsp3	0.138	HF	C5986T	Syn	nsp3
C6286T_Syn_nsp3	0.072	MF	C6286T	Syn	nsp3
A6319G_Syn_nsp3	0.037	LF	A6319G	Syn	nsp3
A6613G_Syn_nsp3	0.035	LF	A6613G	Syn	nsp3
T6954C_I1412T_nsp3	0.138	HF	T6954C	I1412T	nsp3
G8083A_M1788I_nsp3	0.038	LF	G8083A	M1788I	nsp3
C10319T_L89F_nsp5	0.097	LF	C10319T	L89F	nsp5
G11083T_L37F_nsp6	0.037	MF	G11083T	L37F	nsp6
T11288-_Δ106-108_nsp6	0.175	HF	T11288-	Δ106-108	nsp6
C11289-_Δ106-108_nsp6	0.175	HF	C11289-	Δ106-108	nsp6
T11290-_Δ106-108_nsp6	0.175	HF	T11290-	Δ106-108	nsp6
G11291-_Δ106-108_nsp6	0.175	HF	G11291-	Δ106-108	nsp6
G11292-_Δ106-108_nsp6	0.174	HF	G11292-	Δ106-108	nsp6
T11293-_Δ106-108_nsp6	0.174	HF	T11293-	Δ106-108	nsp6
T11294-_Δ106-108_nsp6	0.174	HF	T11294-	Δ106-108	nsp6
T11295-_Δ106-108_nsp6	0.174	HF	T11295-	Δ106-108	nsp6
T11296-_Δ106-108_nsp6	0.174	HF	T11296-	Δ106-108	nsp6
C12053T_L71F_nsp7	0.038	LF	C12053T	L71F	nsp7
C12778T_Syn_nsp9	0.034	LF	C12778T	Syn	nsp9
C13860T_Syn_nsp12	0.036	LF	C13860T	Syn	nsp12
C14120T_P227L_nsp12	0.031	LF	C14120T	P227L	nsp12
C14408T_P323L_nsp12	0.98	HF	C14408T	P323L	nsp12
C14676T_Syn_nsp12	0.139	HF	C14676T	Syn	nsp12

C14805T_Syn_nsp12	0.043	LF	C14805T	Syn	nsp12
C15279T_Syn_nsp12	0.138	HF	C15279T	Syn	nsp12
G15766T_V776L_nsp12	0.032	MF	G15766T	V776L	nsp12
T16176C_Syn_nsp12	0.137	HF	T16176C	Syn	nsp12
G17259T_E341D_nsp13	0.036	LF	G17259T	E341D	nsp13
A17615G_K460R_nsp13	0.034	MF	A17615G	K460R	nsp13
A18424G_N129D_nsp14	0.088	LF	A18424G	N129D	nsp14
C18877T_Syn_nsp14	0.089	MF	C18877T	Syn	nsp14
T19839C_Syn_nsp15	0.036	LF	T19839C	Syn	nsp15
A20268G_Syn_nsp15	0.072	MF	A20268G	Syn	nsp15
G21255C_Syn_nsp16	0.067	MF	G21255C	Syn	nsp16
C21304T_Syn_nsp16	0.082	LF	C21304T	Syn	nsp16
C21614T_L18F_S	0.059	MF	C21614T	L18F	S
C21621A_T20N_S	0.034	LF	C21621A	T20N	S
C21638T_P26S_S	0.037	LF	C21638T	P26S	S
T21765-_Δ69-70_S	0.147	HF	T21765-	Δ69-70	S
A21766-_Δ69-70_S	0.145	HF	A21766-	Δ69-70	S
C21767-_Δ69-70_S	0.145	HF	C21767-	Δ69-70	S
A21768-_Δ69-70_S	0.145	HF	A21768-	Δ69-70	S
T21769-_Δ69-70_S	0.145	HF	T21769-	Δ69-70	S
G21770-_Δ69-70_S	0.145	HF	G21770-	Δ69-70	S
G21974T_D138Y_S	0.046	LF	G21974T	D138Y	S
T21991-_Δ144_S	0.134	HF	T21991-	Δ144	S
T21992-_Δ144_S	0.134	HF	T21992-	Δ144	S
A21993-_Δ144_S	0.135	HF	A21993-	Δ144	S
G22132T_R190S_S	0.033	LF	G22132T	R190S	S
C22227T_A222V_S	0.074	MF	C22227T	A222V	S
C22444T_D294D_S	0.037	LF	C22444T	D294D	S
T22917G_L452R_S	0.054	LF	T22917G	L452R	S
G22992A_S477N_S	0.037	MF	G22992A	S477N	S
G23012A_E484K_S	0.08	MF	G23012A	E484K	S
A23063T_N501Y_S	0.181	HF	A23063T	N501Y	S
C23271A_A570D_S	0.136	HF	C23271A	A570D	S
A23403G_D614G_S	0.989	HF	A23403G	D614G	S
C23525T_H655Y_S	0.037	LF	C23525T	H655Y	S
C23604A_P681H_S	0.167	HF	C23604A	P681H	S
C23604G_P681R_S	0.034	LF	C23604G	P681R	S
C23709T_T716I_S	0.139	HF	C23709T	T716I	S
T24506G_S982A_S	0.137	HF	T24506G	S982A	S
C24642T_T1027I_S	0.035	LF	C24642T	T1027I	S
G24914C_D1118H_S	0.137	HF	G24914C	D1118H	S
G25088T_V1176F_S	0.082	LF	G25088T	V1176F	S
G25563T_Q57H_ORF3a	0.285	MF	G25563T	Q57H	ORF3a
C25710T_L106L_ORF3a	0.031	MF	C25710T	L106L	ORF3a
G25907T_G172V_ORF3a	0.087	LF	G25907T	G172V	ORF3a
T26149C_S253P_ORF3a	0.037	LF	T26149C	S253P	ORF3a

C26681T_F53F_M	0.047	LF	C26681T	F53F	M
C26735T_Y71Y_M	0.08	MF	C26735T	Y71Y	M
C26801G_L93L_M	0.068	MF	C26801G	L93L	M
T27299C_I33T_ORF6	0.04	LF	T27299C	I33T	ORF6
C27944T_H17H_ORF8	0.035	LF	C27944T	H17H	ORF8
C27964T_S24L_ORF8	0.103	LF	C27964T	S24L	ORF8
C27972T_Q27stop_ORF8	0.139	HF	C27972T	Q27stop	ORF8
G28048T_R52I_ORF8	0.136	HF	G28048T	R52I	ORF8
A28095T_K68stop_ORF8	0.05	MF	A28095T	K68stop	ORF8
A28111G_Y73C_ORF8	0.138	HF	A28111G	Y73C	ORF8
G28167A_E92K_ORF8	0.034	LF	G28167A	E92K	ORF8
C28253T_F120F_ORF8	0.043	MF	C28253T	F120F	ORF8
A28271-_---_Intergenic	0.143	HF	A28271-	---	Intergenic
G28280C_D3H_N	0.129	HF	G28280C	D3H	N
A28281T_D3V_N	0.128	HF	A28281T	D3V	N
T28282A_D3E_N	0.129	HF	T28282A	D3E	N
C28472T_P67S_N	0.09	LF	C28472T	P67S	N
C28512G_P80R_N	0.034	LF	C28512G	P80R	N
C28854T_S194L_N	0.098	MF	C28854T	S194L	N
C28869T_P199L_N	0.102	LF	C28869T	P199L	N
A28877T_S202C_N	0.037	LF	A28877T	S202C	N
G28878C_S202T_N	0.036	LF	G28878C	S202T	N
G28881A_R203K_N	0.456	MF	G28881A	R203K	N
G28882A_R203R_N	0.455	MF	G28882A	R203R	N
G28883C_G204R_N	0.457	MF	G28883C	G204R	N
C28887T_T205I_N	0.055	MF	C28887T	T205I	N
C28932T_A220V_N	0.07	MF	C28932T	A220V	N
G28975T_M234I_N	0.046	LF	G28975T	M234I	N
C28977T_S235F_N	0.141	HF	C28977T	S235F	N
T29148C_I292T_N	0.036	LF	T29148C	I292T	N
G29402T_D377Y_N	0.052	LF	G29402T	D377Y	N
G29645T_V30L_ORF10	0.07	MF	G29645T	V30L	ORF10


Non-synonymous nsp
Non-synonymous Structural
Synonymous
a
b
c
d
e
f
g
h
i
Normalized Relative Frequency (NRFp)
C3267T_T183I_nsp3
C5388A_A890N_nsp3
T6954C_I1412T_nsp3
T11288-_Δ106-108_nsp6
C11289-_Δ106-108_nsp6
T11290-_Δ106-108_nsp6
G11291-_Δ106-108_nsp6
G11292-_Δ106-108_nsp6
T11293-_Δ106-108_nsp6
T11294-_Δ106-108_nsp6
T11295-_Δ106-108_nsp6
T11296-_Δ106-108_nsp6
C14408T_P323L_nsp12
T21765-_Δ69-70_S
A21766-_Δ69-70_S
C21767-_Δ69-70_S
A21768-_Δ69-70_S
T21769-_Δ69-70_S
G21770-_Δ69-70_S
T21991-_Δ144_S
T21992-_Δ144_S
A21993-_Δ144_S
A23063T_N501Y_S
C23271A_A570D_S
A23403G_D614G_S
C23604A_P681H_S
C23709T_T716I_S
T24506G_S982A_S
G24914C_D1118H_S
C27972T_Q27stop_ORF8
G28048T_R52I_ORF8
A28111G_Y73C_ORF8
G28280C_D3H_N
A28281T_D3V_N
T28282A_D3E_N
C28977T_S235F_N
C241T_---_5-UTR
C913T_Syn_nsp2
C3037T_Syn_nsp3
C5986T_Syn_nsp3
C14676T_Syn_nsp12
C15279T_Syn_nsp12
T16176C_Syn_nsp12
A28271-_---_Intergenic
C1059T_T85I_nsp2
G11083T_L37F_nsp6
G15766T_V776L_nsp12
A17615G_K460R_nsp13
C21614T_L18F_S
C22227T_A222V_S
G22992A_S477N_S
G23012A_E484K_S
G25563T_Q57H_ORF3a
C25710T_L106L_ORF3a
C26735T_Y71Y_M
C26801G_L93L_M
A28095T_K68stop_ORF8
C28253T_F120F_ORF8
C28854T_S194L_N
G28881A_R203K_N
G28882A_R203R_N
G28883C_G204R_N
C28887T_T205I_N
C28932T_A220V_N
G29645T_V30L_ORF10
T445C_Syn_nsp1
C4543T_Syn_nsp3
C6286T_Syn_nsp3
C18877T_Syn_nsp14
A20268G_Syn_nsp15
G21255C_Syn_nsp16
C3828T_S370L_nsp3
A5648C_K977Q_nsp3
G8083A_M1788I_nsp3
C10319T_L89F_nsp5
C12053T_L71F_nsp7
C14120T_P227L_nsp12
G17259T_E341D_nsp13
A18424G_N129D_nsp14
C21621A_T20N_S
C21638T_P26S_S
G21974T_D138Y_S
G22132T_R190S_S
C22444T_D294D_S
T22917G_L452R_S
C23525T_H655Y_S
C23604G_P681R_S
C24642T_T1027I_S
G25088T_V1176F_S
G25907T_G172V_ORF3a
T26149C_S253P_ORF3a
C26681T_F53F_M
T27299C_I33T_ORF6
C27944T_H17H_ORF8
C27964T_S24L_ORF8
G28167A_E92K_ORF8
C28472T_P67S_N
C28512G_P80R_N
C28869T_P199L_N
A28877T_S202C_N
G28878C_S202T_N
G28975T_M234I_N
T29148C_I292T_N
G29402T_D377Y_N
G210T_---_5-UTR
C313T_Syn_nsp1
T733C_Syn_nsp1
C2749T_Syn_nsp3
A6319G_Syn_nsp3
A6613G_Syn_nsp3
C12778T_Syn_nsp9
C13860T_Syn_nsp12
C14805T_Syn_nsp12
T19839C_Syn_nsp15
C21304T_Syn_nsp16


a
Relative Frequency Change of HF mutations
Relative Level of International Travel Controls
Africa
Asia
Europe
NorthAmerica
Oceania
SouthAmerica
1.00
0.75
0.50
0.25
0.00
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
2021-Jan
2021-Feb
2021-Mar
b
Relative Frequency Change of HF mutations
Relative Number of Daily Flight Departures
c
R = −0.37 , R2 = 0.14 , p = 0.001
Relative Level of International Travel Controls
Relative Frequency Change of HF mutations
d
R = 0.24 , R2 = 0.06 , p = 0.034
Relative Number of Daily Flight Departures
Relative Frequency Change of HF mutations