scientific reports

Check for updates

OPEN

Dynamics of SARS-CoV-2 mutations reveals regional-specificity and similar trends of N501 and high-frequency mutation N501Y in different levels of control measures

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[2,4] & Roberto Pineda Chavarría[1]

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 to date. Surveillance and monitoring of new mutations in the virus' genome are crucial to our understanding of the adaptation of SARS-CoV-2. Moreover, how the temporal dynamics of these mutations is influenced by control measures and non-pharmaceutical interventions (NPIs) is poorly understood. Using 1,058,020 SARS-CoV-2 from sequenced COVID-19 cases from 98 countries (totaling 714 country-month combinations), we perform a normalization by COVID-19 cases to calculate the relative frequency of SARS-CoV-2 mutations and explore their dynamics over time. 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 enable us to examine viral adaptation and evaluate the effects of implemented control measures in virus evolution during the pandemic. We showed that medium-frequency mutations are characterized by high prevalence in specific regions and/or in constant competition with other mutations in several regions. Finally, taking N501Y mutation as representative of high-frequency mutations, we showed that level of control measure stringency negatively correlates with the effective reproduction number of SARS-CoV-2 with high-frequency or not-high-frequency and both follows similar trends in different levels of stringency.

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, significant genomic sequencing efforts have played a central role in furthering our understanding of the evolutionary dynamics of the virus. This has allowed the identification of mutations that appeared early in the pandemic (and that now seem to be fixed in the population[2–6]), as well as monitoring of the effectiveness of vaccines against variants coding for mutations in the spike[7–12]. Both underscore the importance of timely identification and surveillance of mutations with significant representation in the population, to efforts aimed at containing transmission of the virus.

[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. ✉email: santiago.justo@urp.edu.pe

The combination of virus spread by droplets through close contact[13,14], the large number of asymptomatic cases[15,16], the absence of effective pharmaceutical treatments at the beginning of the pandemic and the delays in production and distribution of vaccines[17], 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 (NPIs) and the decrease in the number of cases[18,19], the reproductive number[18,20,21], the case fatality rate[22], the contagion rate[23], and the number of SARS-CoV-2 importations[24,25]. By contrast, the effect of NPIs on specific mutations has been less well-studied. Pachetti et al.[22] analyzed how lockdown policies might have influenced the dynamics of some SARS-CoV-2 mutations; however, results are primarily qualitative and little quantitative description of the reported effect is provided. Muller et al.[26] 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, using 1,058,020 genomes from sequenced COVID-19 cases, we analyze the temporal dynamics of SARS-CoV-2 mutations estimated to be present in more than 3% of global COVID-19 cases. We then investigate whether mutations are region-specific and if there is a correlation between level of lockdown policies and the effective reproduction number of specific mutations.

Results and discussions

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. The relative frequency of cases where a mutation is present was named Normalized Relative Frequency of a genomic position: 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.

This normalization allowed us to identify mutations that have not been reported in other global studies, such as that of Castonguay et al.[27]. This is because in many countries the number of sequenced genomes is low and certain mutations could go unnoticed. Thus, 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.

Table S1 summarizes the features of these 115 mutations. Based on those 115 mutations, we calculated a dN/dS ratio of 4.1 that could imply positive selection occurring in the SARS-CoV-2 genome. Additionally, S and N proteins did not show synonymous mutations and presents ~ 74% of the total non-synonymous mutations suggesting that positive selection is predominantly in those two ORFs.

Mutations show three types of temporal dynamics. The dynamics of the 115 mutations were analyzed through calculating the NRFp in each month from January 2020 to April 2021 (Fig. 1). We assigned type of temporal dynamics to the mutations according to the NRFp in different months and the change of NRFp between months. Thus, three types of temporal dynamics were observed: (i) high-frequency mutations (HF) that never show negative NRFp changes greater than 1%, and increased rapidly in NRFp since their appearance (Fig. 1a), (ii) medium-frequency mutations (MF) that alternates between negative and positive NRFp changes and presents at least one month with NRFp greater than 15% (Fig. 1b), and (iii) low-frequency mutations (LF) that also have an alternation between negative and positive NRFp changes but at a NRFp ever below 15% (Fig. 1c).

HF mutations are characterized by a rapid increase in global frequency following their appearance (Fig. 1a). This could be due to positive selection without competition and/or by other effects related to population dynamics such as control measures implemented by countries aimed at controlling transmission. Mutations in this category appeared in two well-defined stages of the pandemic. The first group is composed of four mutations that now appear to be globally fixed. They emerged at the beginning of the pandemic in January 2020, reaching more than 0.75 NRFp in April 2020 (Fig. 1a, Group 1). The second group rapidly increased in frequency in December 2020, and have continued to increase since then (Fig. 1a, Group 2).

Some HF mutations identified here have been widely reported[28] due to their presence in variants of concern. The first and second groups contained Spike mutations well known due to their possible implications in transmissibility, (e.g. D614G in the first group[29,30] (Fig. S2b)), and vaccine efficacy (e.g. Δ69–70, N501Y, and E484K, all present in the second group[31,32] (Fig. S2b,e)). In the future, analysis of the dynamics of other mutations in this way could help facilitate rapid identification of other mutations of concern.

By contrast, some of the MF and LF mutations that we observed have not been less previously reported to a significant degree, with descriptions either limited to specific countries or regions[33–35], or not reported at all, (e.g. K997Q on nsp3 and S202C on N protein). However, those mutations are present in several months throughout the pandemic and we did not observe evidence of the extinction of any of these mutations (relative frequency of 0 or near to 0 in two or more consecutive months) (Fig. 1b,c).

One possibility for the existence of MF and LF mutations is that some benefits may be conferred to SARS-CoV-2 but competition with other variants prevents rapid increases in their frequency increase across the population. Such dynamics have been observed in evolution experiments for other organisms[36,37]. Furthermore, the coexistence of different lineages of the same organism in the context of frequency-dependent interactions has been reported in yeast[38] and bacteria[39,40], and have highlighted that this can be beneficial for the organism. In the case of virus, epitope diversity and host-specific adaptation can be beneficial for the viral population[41].

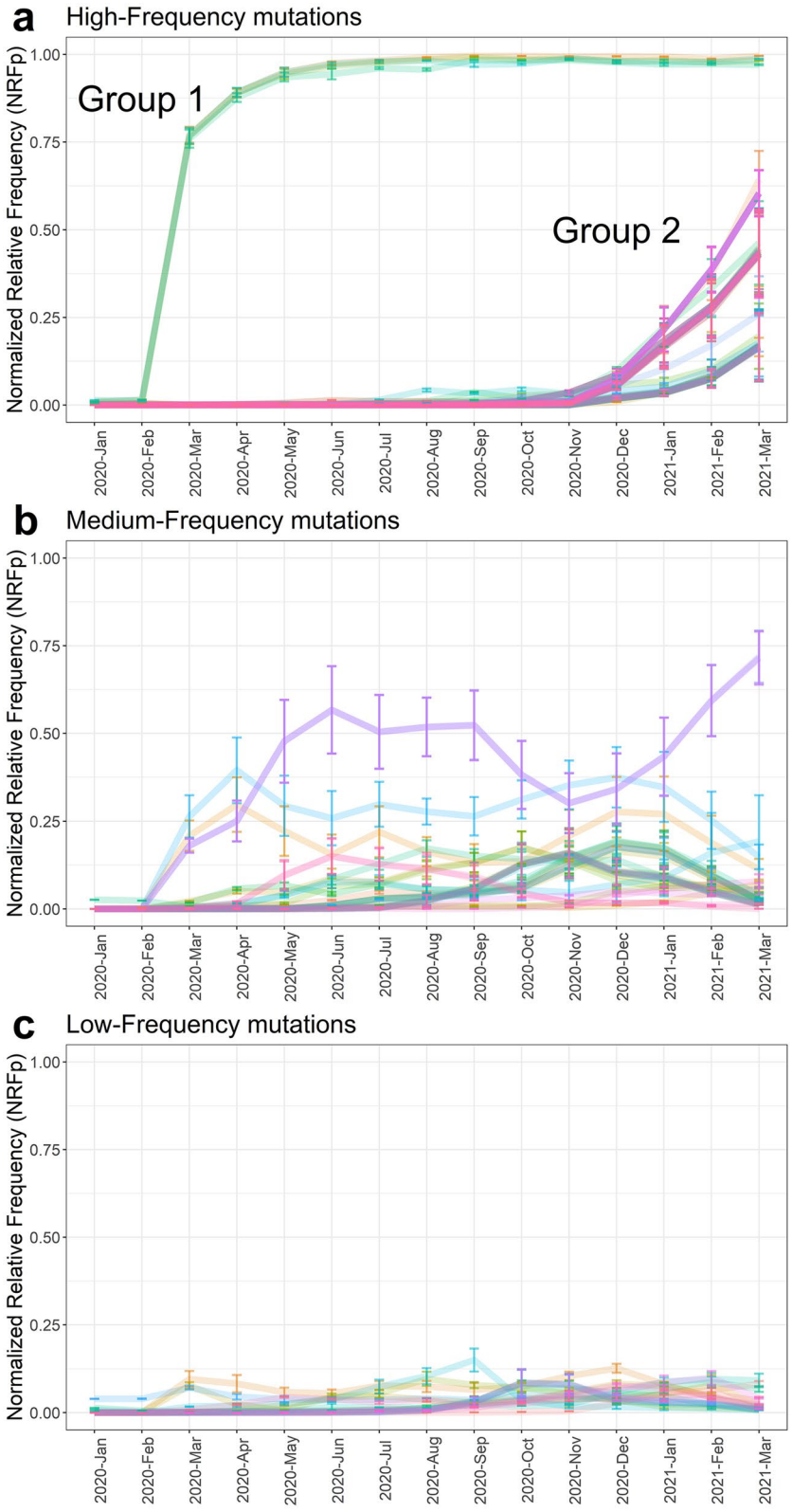

Figure 1. Three different temporal dynamics of SARS-CoV-2 mutations. Normalized by cases Relative Frequency (NRFp) of the mutations by month. **(a)** high-frequency mutations (HF) never show negative NRFp changes greater than 1%, and increased rapidly since their appearance. **(b)** medium-frequency mutations (MF) alternates between negative and positive NRFp changes and presents at least one month with NRFp greater than 15% **(c)** low-frequency mutations (LF) that also have an alternation between negative and positive NRFp changes but at a NRFp ever below 15%. Error bars represent inter-region variation as weighted variance.

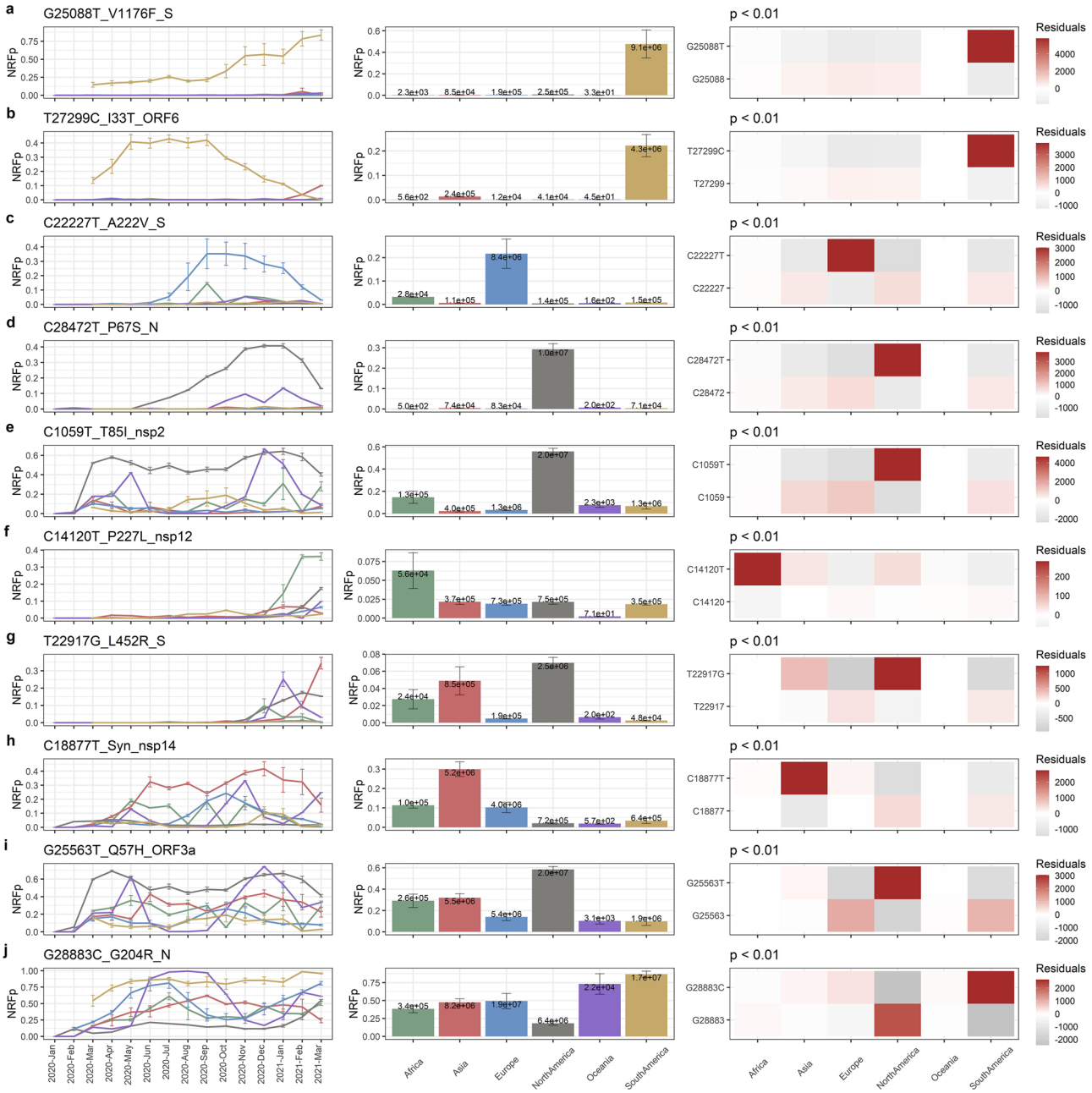

Figure 2. MF mutations are region-specific or have mid-frequencies in several regions. (Left-column) Normalized by cases Relative Frequency (NRFp) of the mutations by month separated by regions (green = Africa, red = Asia, blue = Europe, grey = North America, purple = Oceania, yellow = South America). (Middle-column) Total NRFp by region of the analyzed medium-frequency mutations. Numbers in each bar represents the estimated total number of cases of the particular mutation in that region. (Right-column) Chi-square p-value and Pearson residuals analysis of medium-frequency mutations. Upper line corresponds to the mutant state and the bottom line to the not-mutant state. Grey and red boxes mean negative or positive association with the state, respectively. Intensity of the colors means higher residuals that means greater contribution. **(a–e)** Region-specific medium-frequency mutations and **(f–j)** not-region-specific medium-frequency mutations.

Some of the MF mutations are region-specific while other have medium frequencies in various regions. In our previous work[42], we observed that the mutation T85I in nsp2 has a higher frequency in North America than in other continents. Here, we show that this MF mutation maintains this tendency, persisting since its appearance at a global NRFp of ~ 0.2 (Fig. S3a). Interestingly, and in contrast to HF mutations (that are typically similarly frequent across several analyzed regions, with an exception being a group of recent mutations that are more frequent in South America (Figs. S4, S5)), most of the MF mutations (18 of 29) analyzed here are most frequent in a specific region (Fig. 2).

To explore whether MF mutations showed a region-specific pattern, we analyzed the dynamics of ten subtypes of MF dynamics in six different regions (Africa, Asia, Europe, North America, Oceania, and South America) (Fig. S6). Our results show that five subtypes had a NRFp greater than 0.3 for at least three consecutive months in only one region (Fig. 2a–e, left column). Relatedly, mutations belonging to these subtypes had a higher relative number of cumulative cases (NRFp) in a specific region, compared to other regions (Fig. 2a–e, middle column).

Then, we examined whether the proportions of estimated COVID-19 cases caused by MF mutations were different between regions. Chi-square p-values showed that in all the MF subtypes at least one region have different proportions (Fig. 2, right column). Pearson residuals analysis showed which of the regions have larger or smaller mutant proportion than expected (meaning positive or negative association, red and grey squares, respectively) and which region has a greater degree of association (color intensity). The five subtypes that showed region-specific patterns also showed that just one region is positively associated and that it has the highest degree of association to that specific mutation (Fig. 2a–e, right column). By contrast, other five subtypes showed positive association to more than one region with a variety of degrees of association (Fig. 2f–j, right column).

We further analyze the five subtypes that showed region-specific pattern (Fig. 2a–e). Country analysis of the relative frequencies (Figs. S7, S8) and the cumulative number of cases (Figs. S9, S10) showed that those mutations are found in more than one country of the region. Some of them follows a similar pattern of frequency changes in two or more countries within the region (S7b, S7d and S8), whereas others have a particular pattern of frequency change in one particular country (S7a and S7c). Analysis of the cumulative number of cases by country showed that, although several countries present COVID-19 cases of the particular mutations, in most cases few countries contributes to most cases (S9a, Brazil; S9b, Argentina, Brazil, Chile; S9c, USA; S9d, Canada, Mexico, USA; S10, Italy, Spain, UK).

A decline of the frequency can be seen for some MF mutations in the last months (Fig. 2b–d), this can be explained because new mutations leave out of competition those mutations, or due to a delay between the collection date and the submission date of genomic samples. Using genomic data from August 10th 2021, we reanalyzed three mutations that clearly showed this decline (Fig. 2b (I33T), c (A222V), and d (P67S)). We found very similar patterns in the countries analyzed (Fig. S11, S12), therefore, leave out of competition by other mutations is a more plausible scenario.

LF mutations followed similar patterns to those observed for MF mutations (Figs. S13, S14). Thus, the MF and LF dynamics seems to be due to: (i) high prevalence of mutations in specific regions, (ii) globally dispersed beneficial mutations in constant competition with other variants, or (iii) a combination of these two effects.

SARS-CoV-2 carrying HF_{N501Y} mutation follows similar trends than SARS-CoV-2 without HF_{N501Y} in different levels of control measures. 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) lead us to hypothesize that, at least part of this abrupt increase is due to the fact that limited or minimal levels of control measures and NPIs may permit that HF mutations to spread even faster than not-HF mutations that when stronger control measures and stronger NPIs are present. An alternative hypothesis could be that strict control measures give a large competitive advantage to more transmissible variants (HF mutations), enabling them to persist and continuing to transmit, whilst their less transmissible counterparts (not-HF mutations) die out.

To test these hypotheses, we analyzed whether different degree of control measures could affect differently to SARS-CoV-2 genomes bearing the HF mutation N501Y (HF_{N501Y}) or not bearing the HF mutation N501Y (not-HF_{N501Y}) in nine countries that have more than 15 sequenced genomes per week during March 2020 to April 2021. We selected this mutation because it is present in three variants of concern (B.1.1.7, B.1.35, and P.1)[43] and is a good example of the behavior of HF mutations (Fig. 1a, Supplementary Fig. S2a). Additionally, and in contrast to HF mutations belonging to group 2, mutations in the first group of HF mutations (Fig. 1a, group 1) may have been aided by founder effects in the early stages of the pandemic. For this reason, we did not analyze them in this part of our study.

First, we estimated the effective reproduction number (Rt) of HF_{N501Y} or not-HF_{N501Y} (Fig. 3a) and measure the correlation with the level of stringency (Fig. 3b). The level of stringency is a measure of the level of control policies based on nine response indicators including school closing, workplace closing, cancel public events, restrictions on gathering size, close public transport, stay-at-home requirements, restrictions on internal movement, restriction on international travel and public information campaigns[44].

We found significant negative correlation between the Rt after 14 days that the level of stringency was implemented and the level of stringency in eight of the nine countries analyzed (Fig. 3b). In all these eight countries linear regression model explained at least 23% of the variance in the Rt of HF_{N501Y} (Fig. 3b), and the effect size measured by the R-value of spearman correlation showed in the worst case a value of 0.48, with all the others R-value between 0.5 and 0.81 (Fig. 3b). In the case of India, the Rt of HF_{N501Y} showed a positive correlation with level of stringency. It is known that efforts in molecular testing in India have changed during the pandemic[45] Time-varying differences in the intensity and capacity of molecular testing can produce significant biases in the estimation of Rt. Overall however, our results show a significant negative correlation between degree of control measure stringency and Rt in eight of the nine countries analyzed.

We also found that, independently of the level of stringency imposed, the Rt of HF_{N501Y} was significantly higher than not-HF_{N501Y}, potentially explaining why HF_{N501Y} increase its frequency faster than not-HF_{N501Y} since its appearance in the nine countries considered here (Fig. 4a). Interestingly, when we analyzed the Rt of SARS-CoV-2 genomes bearing an MF mutation (MF_{R203K}) and compare it with the Rt of genomes without the MF mutation (not-MF_{R203K}) we observed that in some stages of the pandemic the Rt of MF_{R203K} is higher than not-MF_{R203K} but in other cases the opposite was observed (Fig. S15). This explains why this mutation did not increases its frequency steadily and can be an evidence of constant competition between MF_{R203K} and not-MF_{R203K}.

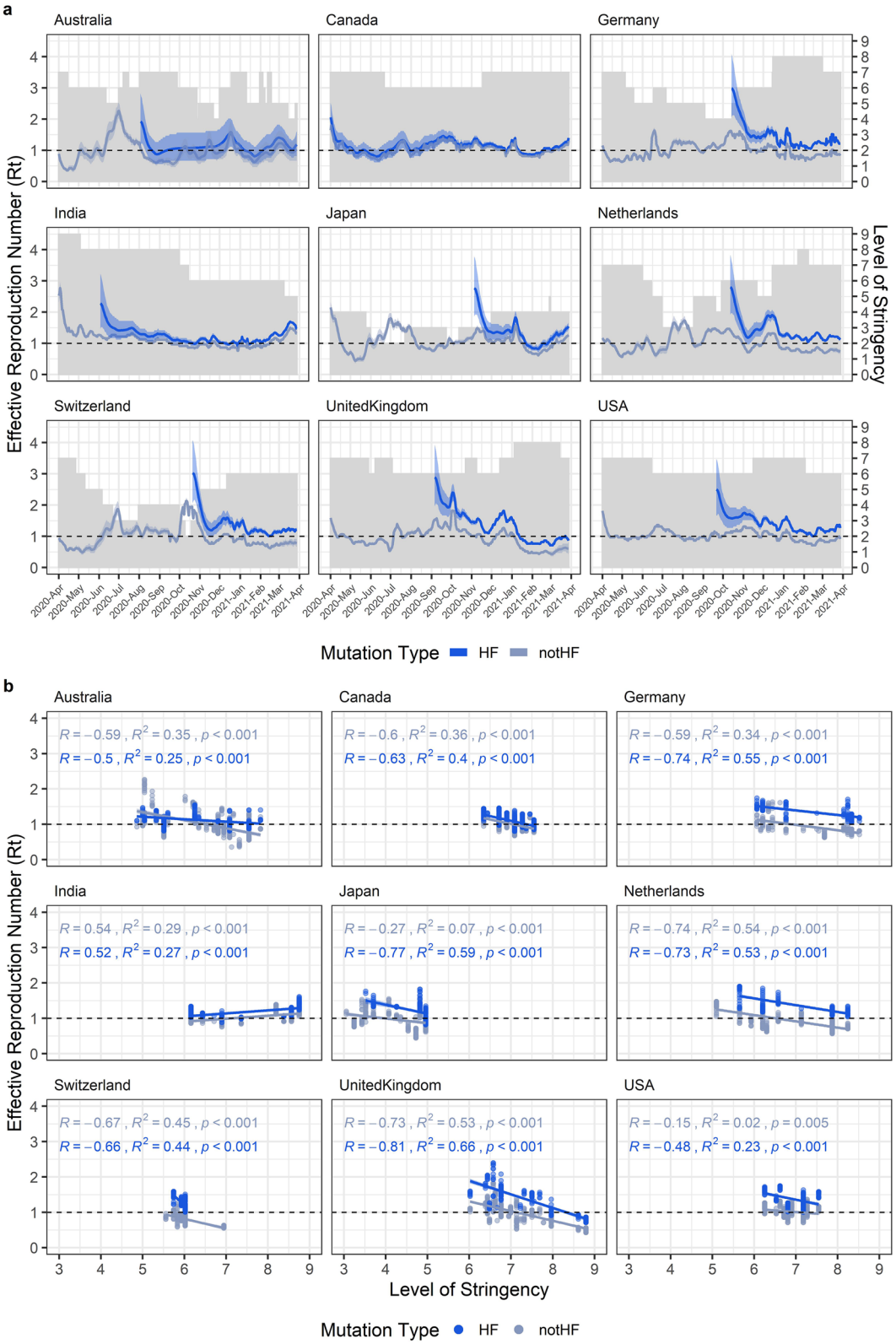

Figure 3. Effective reproduction number (Rt) of HF_{N501Y} and not-HFN_{501Y} are correlated with level of stringency. **(a)** Each panel shows the estimated effective reproduction number of SARS-CoV-2 bearing (blue) or not (grey) the HF mutation N501Y (HF or notHF, respectively) in different countries. Grey bars are showing the level of stringency. Shades show a 97.5% confidence interval in the estimation of Rt. **(b)** Correlation of Rt after 14 days of the implementation of the level of stringency with the level of stringency. Each panel shows the independent analysis of different countries. Spearman correlation values (R), R-square of the linear regression model (R^2), and p-value of the correlation is showed in the left-up of each panel in this order. Colors represent the same as in **(a)**.

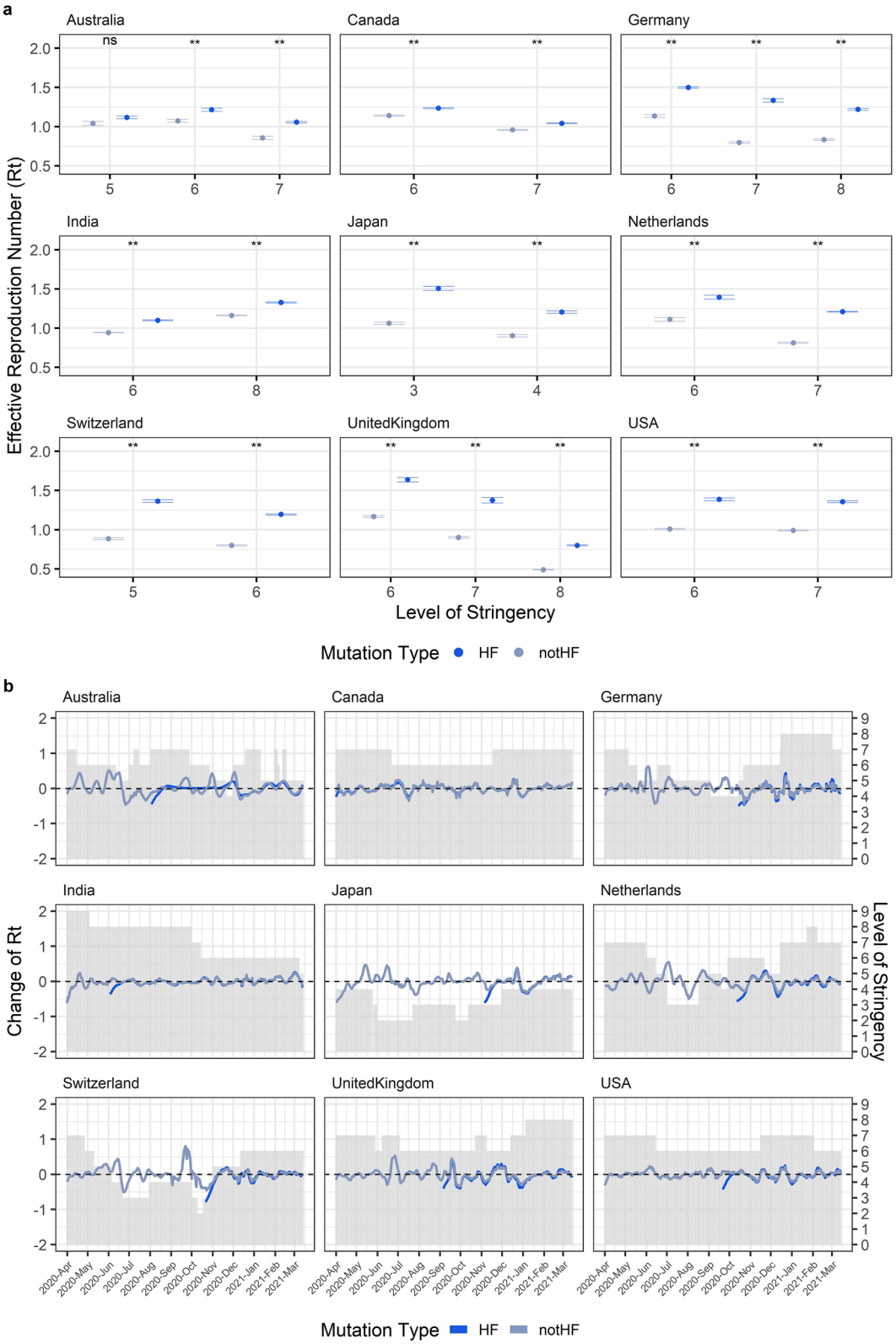

Figure 4. Effective reproduction number (Rt) of HF_{N501Y} is higher than not-HF_{N501Y} but similarly affected by the level of stringency. **(a)** Statistical comparison between the bootstrap distribution of the Rt of SARS-CoV-2 bearing (blue) or not (grey) the HF mutation N501Y (HF or notHF, respectively) in different levels of stringency. Points represent the mean and the lines represent the 25 and 75 percentiles of the bootstrap distribution. **Means p-value lesser than 0.05 and ns means p-value higher than 0.05. **(b)** Plot of the change of Rt in the time. Change of Rt was calculated as the Rt 14 days after the day of interest subtracted to the Rt mean between the day of interest and 13 days after that day. Grey bars are showing the level of stringency. Colors represent the same as in **(a)**.

Finally, to explore whether different stringency levels differentially affect the dynamics and transmission of HF_{N501Y} and not-HF_{N501Y}, we calculated the change in Rt during several months where different levels of stringency were implemented (Fig. 4b). The patterns of the change of Rt between HF_{N501Y} and not-HF_{N501Y} were almost identical (Fig. 4b) and R values of spearman correlation of HF_{N501Y} and not-HF_{N501Y} with Rt were similar in most cases (Fig. 3b), indicating that both could be similarly affected by the changes in stringency levels.

Taken together, although HF_{N501Y} presented higher Rt in lower levels of stringency indicating that HF_{N501Y} spread was likely helped by mild lockdown policies in some stages of the pandemic, this effect was also observed in not-HF_{N501Y}. In conclusion, the results of this section showed control measures and their associated stringency probably affecting HF_{N501Y} and not-HF_{N501Y} in a similar fashion; thus, our two initial hypotheses are not supported by these results. Instead, the rapid increase of frequency of HF_{N501Y} is justified primarily by its generally increased transmissibility (i.e. a higher Rt which is always greater than the Rt of not-HF_{N501Y}), rather than the implementation of specific control measures.

Limitations of the study. Stringency level is calculated from set of policies applied in each country that do not necessarily operate or function the same in different countries due to, for instance, variations in sociocultural and economic factors. Thus, comparisons at country level have variation that limit the reliability and interpretability of the results presented here, especially when compared with other countries. Moreover, different combinations of policies can generate the same level of stringency—the fact that several policies were applied together to generate a stringency index precludes efforts to evaluate the effect of a specific policy on the effective reproduction number of SARS-CoV-2.

After control measures are implemented (reflected as an increase to the stringency index) Rt changes from a higher value to a lower value. This process generates a time-window of intermediate Rt before the Rt reach a plateau that indicates how much the policy lowered the Rt. These intermediate values of Rt introduce a bias in the correlation between Rt and the level of stringency. Furthermore, if a country changes the stringency level in time-windows less than those necessary for the Rt to stabilize, the estimations of correlation get more complicated.

Our correlation analysis showed that in seven of the nine countries analyzed lower levels of stringency are correlated with higher Rt values. This could be an evidence of a possible effect of lockdown policies in the Rt. However, causal inference model is known to be a more accurate approach to test causality.

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. Also, the number of cases is still subjected to bias due to for instance, the difference in the number of tests that each country performs, as occurs in India.

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 partially mitigates differences in available genomes, 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 are described here and named "high-frequency" (HF), "medium-frequency" (MF), and "low-frequency" (LF). The three types are represented in all the months analyzed, and found in non-structural and structural proteins, and synonymous and non-synonymous mutations. 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.

Medium and low-frequency mutations maintain roughly constants global frequency due to their higher prevalence on specific regions and/or because they are in constant competition with other mutations in several regions. We showed some mutations with a high degree of region-specificity and others that presented mid-frequencies in several regions. Higher prevalence in specific regions may be due to specific-host characteristics. Constant competition in several regions may be due to the fact that they are beneficial mutation in the presence of other mutations with a similar degree of benefit. Some mutation can be leave out of competition when others beneficial mutations appear. Our analysis, also shows evidence that some MF mutations have a reduced relative frequency after several months of high frequencies in a specific region.

In this pandemic, human behavior has strongly affected the adaptive process of the SARS-CoV-2 through continuous implementations and changes to implemented control measures. Our analysis presents evidence that the high-frequency mutation N501Y is more transmissible (showed for its greater effective reproduction number) than not-N501Y, but also that control measures do not significantly favor the growth of any one in particular. Instead, we observe that policies have a similar impact on both.

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.[42], 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 29,674 using ViralMSA.py[46,47]. 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 gener-

ated 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. Normalized relative frequencies (NRFp) were calculated as the weighted mean of the relative frequencies in each subalignment with the number of cases as the weight. 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). The NRFp is an estimation of the percentage of global COVID-19 cases where a particular mutation is present. The same procedure was done to obtain the NRFp of the mutations by months or by regions. Data manipulation was done using R and python scripts.

Analysis of region-specific mutations. The frequencies by country-months of each mutation were obtained from the previous calculation. Then, the Normalized relative frequencies (NRFp) by region (Africa, Asia, Europe, North America, Oceania, South America) were calculated as the weighted mean of the relative frequencies of each country-month belonging to a specific region using the number of cases as the weight. Number of cases with a particular mutation in each country was estimated by multiplying the relative frequency of the mutation with the number of cases in a specific country-month. Then, we added the cases belonging to a specific region and chi-square analyses were done using R software[48].

Estimation of effective reproduction number of SARS-CoV-2 mutations. We select nine countries (Australia, Canada, Germany, India, Japan, Netherlands, Switzerland, United Kingdom, USA) with at least 15 sequenced genomes by week from March 2020 to March 2021. Raw number of cases by days were obtained from[49] and used to estimate the number of cases by day for a specific mutation. In the case of MF mutation R203K, R203, and N501, we multiply the relative frequencies of the genomes with the state of interest (R203K, R203 or N501) in a determined week by the number of cases in the day. For instance, if 1 week presented 30% of genomes with the mutation R203K, and the number of cases on Monday of that week was 100. Thus, the estimated number of cases with this mutation in that day was 30. In the case of the HF mutation N501Y we first calculated the relative frequencies of that mutation in each week and then adjusted the relative frequencies to a logistic regression model using R software[50]. The number of cases estimated for the MF and HF mutations by day were used to estimate the effective reproduction number using EpiFilter[51].

Correlation analysis between stringency levels and effective reproduction number. The stringency index by country by day was obtained from[49]. Analysis of Spearman correlations and linear regression models of the effective reproduction number 14 days after the level of stringency was implemented with stringency index in each country by each state (mutant or not mutant) was done using R[48] and the packages ggplot2[52] and ggpubr[50].

Statistical differences between effective reproduction number of SARS-CoV-2 mutations in different levels of stringency. To determine if SARS-CoV-2 with HF_{N501Y} and not-HF_{N501Y} mutations presented statistical differences in Rt in different levels of stringency, we categorize the stringency index in ten levels: 0–10 = 0, 11–20 = 1, 21–30 = 2, 31–40 = 3, 41–50 = 4, 51–60 = 5, 61–70 = 6, 71–80 = 7, 81–90 = 8, and 91–100 = 9. We estimated the distribution of the effective reproduction number 14 days after the level of stringency was implemented in each level of stringency by bootstrap using 1000 replicates. Level of stringency with at least 10 Rt points were considered in the bootstrap analysis. We also used bootstrap methods to estimate the distribution of the difference of the Rt assuming that both Rt (HF_{N501Y} and not-HF_{N501Y}) comes from the same distribution and calculate the p-value of the observed difference.

Calculation of change in time of the effective reproduction number of SARS-CoV-2 mutations. Change of Rt was calculated by subtracting the value of Rt 14 days after the day of interest with the mean of the Rt from the day of interest to 13 days after the day of interest.

Data availability

Publicly available datasets were analyzed in this study. This data can be found at: gisaid.org. All the code used to perform the analysis of this manuscript is publicly available in: https://github.com/sanjusare/Justo_et_al_2021_SR.

Received: 30 November 2020; Accepted: 24 August 2021
Published online: 07 September 2021

References

1. World Health Organization (2021). https://covid19.who.int/ (Accessed 16 May 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. https://doi.org/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. *BioRxiv.* 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. https://doi.org/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.* https://doi.org/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. *MedRxiv.* https://doi.org/10.1101/2020.12.21.20248640v1 (2020).
9. Jangra, S. *et al.* SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe.* https://doi.org/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.* https://doi.org/10.2807/1560-7917.ES.2020.26.1.2002106 (2021).
11. Liu, Y. *et al.* The N501Y spike substitution enhances SARS-CoV-2 transmission. *BioRxiv.* 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. *BioRxiv.* https://doi.org/10.1101/2020.12.14.422555v6 (2021).
13. Santarpia, J. *et al.* Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care. *Sci. Rep.* **10**, 12732 (2020).
14. Leung, N. *et al.* Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nat. Med.* **26**, 676–680 (2020).
15. Sayampanathan, A. *et al.* Infectivity of asymptomatic versus symptomatic COVID-19. *The Lancet* **397**(10269), 93–94 (2021).
16. Alene, M. *et al.* Magnitud of asymptomatic COVID-19 cases throughout the course of infection: A systematic review and meta-analysis. *PLoS ONE* **16**(3), e0249090. https://doi.org/10.1371/journal.pone.0249090 (2021).
17. Kim, J., Marks, F. & Clemens, J. Looking beyond COVID-19 vaccine phase 3 trials. *Nat. Med.* **27**, 205–211 (2021).
18. Hyafil, A. & Moriña, D. Analysis of the impact of lockdown on the reproduction number of the SARS-CoV-2 in Spain. *Gac. Sanit.* https://doi.org/10.1016/j.gaceta.2020.05.003 (2020).
19. Hsiang, S. *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**(7820), 262–267 (2020).
20. 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).
21. 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).
22. 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).
23. Larrosa, J. M. SARS-CoV-2 in Argentina: Lockdown, mobility, and contagion. *J. Med. Virol.* **93**(4), 2252–2261 (2021).
24. 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. N. Z. J. Public Health* **44**(4), 257–259. https://doi.org/10.1111/1753-6405.13016 (2020).
25. 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).
26. Muller, N. F. *et al.* Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Sci. Transl. Med.* https://doi.org/10.1126/scitranslmed.abf0202 (2021).
27. Castonguay, N., Zhang, W. & Langlois, M. Meta-analysis and structural dynamics of the emergence of genetic variants of SARS-CoV-2. *MedRxiv.* https://doi.org/10.1101/2021.03.06.21252994v2 (2021).
28. Plante, J. *et al.* The variant gambit: COVID-19's next move. *Cell Host Microbe* **29**, 508 (2021).
29. Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592**, 122–127 (2021).
30. Hou, Y. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).
31. Supasa, P. *et al.* Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell* **184**, 2201–2211 (2021).
32. Dejnirattisai, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939–2954 (2021).
33. Hodcroft, E. E. *et al.* Emergence of multiple lineages of SARS-CoV-2 spike protein variants affecting amino acid position 677. *MedRxiv.* https://doi.org/10.1101/2021.02.12.21251658v3 (2020).
34. Nagy, Á., Pongor, S. & Győrffy, B. Different mutations in SARS-CoV-2 associate with severe and mild outcome. *Int. J. Antimicrob. Agents* **57**, 106272. https://doi.org/10.1016/j.ijantimicag.2020.106272 (2021).
35. Zrelovs, N. *et al.* First report on the latvian SARS-CoV-2 isolate genetic diversity. *Front. Med.* **8**, 626000. https://doi.org/10.3389/fmed.2021.626000 (2021).
36. Good, B., McDonald, M., Barrick, J., Lenski, R. & Desai, M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
37. Luksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).
38. Frenkel, E. *et al.* Crowded growth leads to the spontaneous evolution of semistable coexistence in laboratory yeast populations. *PNAS* **112**(36), 11306–11311 (2015).
39. Maddamsetti, R., Lenski, R. & Barrick, J. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* **200**, 619–631 (2015).
40. Rozen, D. & Lenski, R. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *Am. Nat.* **155**(1), 24–35 (2000).
41. Parameswaran, P. *et al.* Intrahost selection pressures drive rapid dengue virus microevolution in acute human infections. *Cell Host Microbe* **22**, 400–410 (2017).
42. Justo, S. *et al.* Global geographic and temporal analysis of SARS-CoV-2 haplotypes normalized by COVID-19 cases during the pandemic. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2021.612432 (2021).
43. Konings, F. *et al.* SARS-CoV-2 Variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* **6**, 821–823 (2021).
44. Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538 (2021).
45. Unnikrishnan, J., Mangalathu, S. & Kutty, R. Estimating under-reporting of COVID-19 cases in Indian states: An approach using a delay-adjusted case fatality ratio. *BMJ Open* **11**, e042584 (2021).
46. Moshiri, N. ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes. *Bioinformatics* **37**(5), 714–716 (2020).
47. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).
48. Core Team (2021). *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021). http://www.R-project.org/ (Accessed 16 May 2021).
49. Ritchie, H., *et al. Coronavirus Pandemic (COVID-19)* (2020). https://ourworldindata.org/coronavirus (Accessed 16 May 2021).
50. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots* (2020). https://CRAN.R-project.org/package=ggpubr (Accessed 16 May 2021).

51. Parag, K. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. *MedRxiv.* https://doi.org/10.1101/2020.09.14.20194589v3 (2021).
52. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016) (Accessed 25 July 2021).

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. We thank Prof. Jose Luis Mena (Faculty of Biological Sciences—Universidad Ricardo Palma) and PhD(c). Carlos Prete (Escola Politecnica—University of Sao Paulo) for their valuable help in the epidemiological and statistical analysis. We thank Prof. Shaker Chuck Farah (Institute of Chemistry—University of Sao Paulo) and PhD(c). Charlie Whittaker (Faculty of Medicine—Imperial College London) for English writing corrections and helpful comments. 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.

Author contributions
S.J.A. designed the study. S.J.A., D.Z.S., C.H.R., G.L.B., A.C.C., R.G.-S.C., C.S.U.C. and R.P.C. analyzed the data. S.J.A., C.S.U.C. and G.U.-C. wrote python and R scripts. The manuscript was written by S.J.A., D.Z.S., C.H.R., G.L.B., A.C.C., R.G.-S.C. and C.S.U.C. All authors discussed the methodologies, results, and read and approved the manuscript.

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.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97267-7.

Correspondence and requests for materials should be addressed to S.J.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.