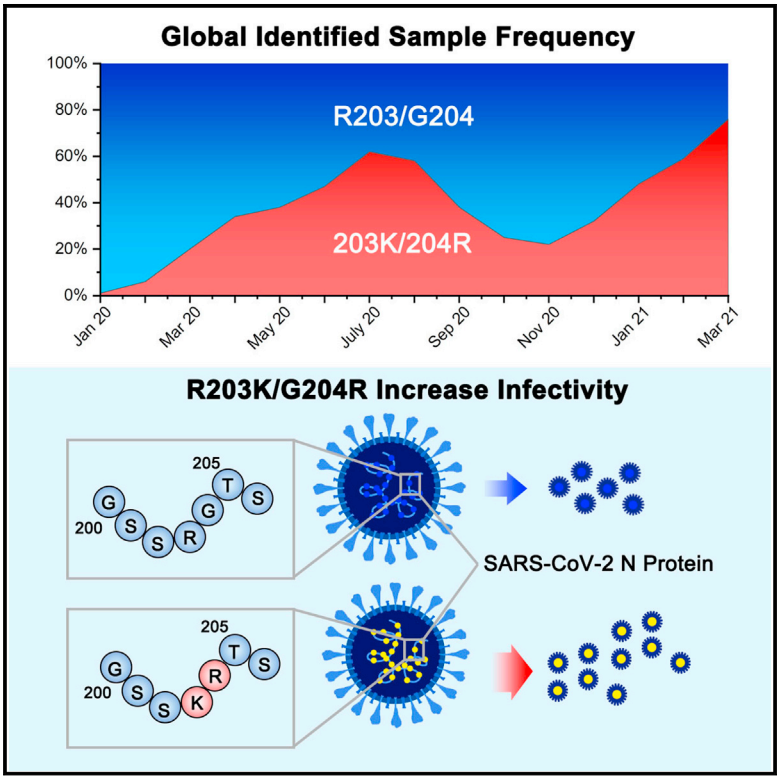
Article

Cell Host & Microbe

Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2

Graphical abstract



Authors

Haibo Wu, Na Xing, Kaiwen Meng, ..., Xiaoyuan Lin, Geng Meng, Zhenglin Zhu

Correspondence

lxy2019@cqu.edu.cn (X.L.), mg@cau.edu.cn (G.M.), zhuzl@cqu.edu.cn (Z.Z.)

In brief

Wu et al. demonstrate that nucleocapsid mutations R203K/G204R in SARS-CoV-2 confer a replication advantage over preceding variants and increase virus fitness and virulence in hamsters. R203K/G204R incur positive selection and associate with the emergence of B.1.1.7 (Alpha). This study highlights the importance of nucleocapsid mutations in SARS-CoV-2 evolution.

Highlights

- SARS-CoV-2 nucleocapsid mutations R203K/G204R associate with B.1.1.7 (Alpha) emergence
- R203K/G204R variants possess a replication advantage over the preceding lineages
- R203K/G204R variants show enhanced infectivity and disease severity in the hamster model

Wu et al., 2021, Cell Host & Microbe *29*, 1788–1801
December 8, 2021
https://doi.org/10.1016/j.chom.2021.11.005

CellPress

CellPress
OPEN ACCESS

Cell Host & Microbe

Article

Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2

Haibo Wu,[1,6] Na Xing,[2,6] Kaiwen Meng,[3,6] Beibei Fu,[1,6] Weiwei Xue,[4] Pan Dong,[1] Wanyan Tang,[5] Yang Xiao,[1] Gexin Liu,[1] Haitao Luo,[1] Wenzhuang Zhu,[3] Xiaoyuan Lin,[1,*] Geng Meng,[3,*] and Zhenglin Zhu[1,7,*]

[1]School of Life Sciences, Chongqing University, No. 55 Daxuecheng South Road, Shapingba, Chongqing 401331, China
[2]Institute of Virology, Free University of Berlin, Robert-von-Ostertag-Str. 7-13, Berlin 14163, Germany
[3]College of Veterinary Medicine, China Agricultural University, No. 2 Yuanmingyuan West Road, Beijing 100094, China
[4]School of Pharmaceutical Sciences, Chongqing University, No. 55 Daxuecheng South Road, Shapingba, Chongqing 401331, China
[5]Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital, No. 181 Hanyu Road, Shapingba, Chongqing 400030, China
[6]These authors contributed equally
[7]Lead contact
*Correspondence: lxy2019@cqu.edu.cn (X.L.), mg@cau.edu.cn (G.M.), zhuzl@cqu.edu.cn (Z.Z.)
https://doi.org/10.1016/j.chom.2021.11.005

SUMMARY

Previous work found that the co-occurring mutations R203K/G204R on the SARS-CoV-2 nucleocapsid (N) protein are increasing in frequency among emerging variants of concern or interest. Through a combination of *in silico* analyses, this study demonstrates that R203K/G204R are adaptive, while large-scale phylogenetic analyses indicate that R203K/G204R associate with the emergence of the high-transmissibility SARS-CoV-2 lineage B.1.1.7. Competition experiments suggest that the 203K/204R variants possess a replication advantage over the preceding R203/G204 variants, possibly related to ribonucleocapsid (RNP) assembly. Moreover, the 203K/204R virus shows increased infectivity in human lung cells and hamsters. Accordingly, we observe a positive association between increased COVID-19 severity and sample frequency of 203K/204R. Our work suggests that the 203K/204R mutations contribute to the increased transmission and virulence of select SARS-CoV-2 variants. In addition to mutations in the spike protein, mutations in the nucleocapsid protein are important for viral spreading during the pandemic.

INTRODUCTION

Since the global outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) began in 2019, COVID-19 has caused more than 246 million confirmed infections and more than 5 million deaths worldwide. SARS-CoV-2 is a rapidly evolving RNA virus that causes severe pneumonia. The virus threatens the lives of individuals in older age groups and those with chronic conditions. Due to its genetic proofreading mechanism (Smith et al., 2013), SARS-CoV-2 (Wu et al., 2020) does not have a high mutation rate (Mercatelli and Giorgi, 2020). However, its short generation time and large population size have enabled this virus to evolve rapidly and mutate continually during transmission. The resulting genomic variations may contribute to the severity of disease and the efficiency of transmission.

In general, most viral mutations are deleterious to the virus and thus disappear quickly, whereas mutations that are selectively neutral or advantageous to viral fitness may persist and increase in frequency. However, it is challenging to classify a mutation as either selectively neutral or under positive selection. In a newly emerging virus such as SARS-CoV-2 in particular, a new mutation with an increasing frequency may result from neutral epidemiological processes such as genetic bottlenecks following founder events and range expansion.

The virus spike glycoprotein (S) directly interacts with the human ACE2 protein (Hussain et al., 2020) and is the target of vaccines and therapeutics (Salvatori et al., 2020). Thus, previous research has focused on adaptive SARS-CoV-2 mutants in S, such as D614G, N501Y (Cheng et al., 2021), and E484K (Jangra et al., 2021). D614G mutants show a dramatically increased identified sample frequency (IF) and increases in fitness, infectivity and fatality (Korber et al., 2020; Mok et al., 2020; Plante et al., 2021; Rochman et al., 2021; Trucchi et al., 2021; Zhu et al., 2021b). N501Y increases virus infectivity (Zhao et al., 2021a) and confers high resistance to neutralization (Garcia-Beltran et al., 2021). E484K has the ability to evade neutralization by most monoclonal antibodies (Hoffmann et al., 2021; Wang et al., 2021). In addition to these adaptive S protein mutations, there should be adaptive mutations in other viral components that distinguish these lineages from the original virus. These mutations may also functionally contribute to the virulence of virus.

The identification and evolutionary analyses of these mutations are important.

R203K/G204R are co-occurring mutations in the N protein (another structural protein of the virion) that are rapidly increasing in frequency and show a potential association with the infectivity of the virus (Zhu et al., 2021b). These mutations are carried by the increasingly frequent lineages B.1.1.7 (Alpha) (Collier et al., 2021; C. Caserta et al., 2021) and P.1 (Gamma) (Dejnirattisai et al., 2021; Faria et al., 2021). The selection signatures of these mutations were identified in our previous work (Zhu et al., 2021b). We continued tracking the evolution of R203K/G204R based on all documented SARS-CoV-2 genome sequences on a monthly basis and found that these mutations showed a second period of rapid expansion accompanying the emergence of B.1.1.7. The R203K/G204R mutations are becoming dominant in the worldwide pandemic and may have positive effects on the fitness of SARS-CoV-2. Thus, a thorough evaluation of the evolutionary and functional effects of R203K/G204R is important for understanding the effects of N mutations and the contribution of N mutations to rapidly increasing lineages. Hence, we constructed an R203K/G204R mutant virus by site-directed mutagenesis. Through experimental investigations conducted in cell lines, hamsters, and a human airway tissue model, we identified and validated increases in the infectivity and fitness of 203K/204R variants. In hamster lung tissues, we observed an increased severity of 203K/204R virus infection, as was implied by epidemic surveys and global clinical data statistics. On the basis of experiments and structural prediction, we concluded that it is possible that a change in the N protein charge resulted in enhanced virus replication and ultimately increased infectivity and fitness (Figure 1). These results indicated that the R203K/G204R N mutations, which may act in coordination with N501Y, are associated with the increased transmission (Washington et al., 2021) and virulence (Davies et al., 2021) of B.1.1.7 and P.1. Thus, the N protein R203K/G204R mutations deserve more attention during SARS-CoV-2 surveillance in the future.

RESULTS

The rapid spread of R203K/G204R mutant viruses worldwide

We performed population genomic analyses of 884,736 full-length SARS-CoV-2 genomes (collected from December 2019 to March 2021; "GList_2103.xls" in Data S1) documented in GISAID (Shu and McCauley, 2017) to track the changes in SARS-CoV-2 mutations. We found 96 mutations with a monthly IF higher than 0.05 ("IF_96.pdf" in Data S1). Through pairwise linkage disequilibrium analyses of these 96 mutation sites (Tables S1A and S1B and Figure S1A), we identified 12 mutation linkage groups (LGs), designated LG_1 to LG_12 hereafter for convenience. There were 26 singleton mutations without any association with an LG. LG_3 contained three adjacent nucleotide mutations, at 28,881 to 28,883, that were completely linked ($\rho^2 > 0.99$, Table S1B). The corresponding amino acid changes at the protein level are co-occurring R203K/G204R substitutions in the N protein. We found that R203K/G204R have rapidly spread worldwide. The global IF of R203K/G204R increased from nearly zero in January 2020 to more than 70% in March 2021 (Figures 2A and 2B). We tracked the IFs of 203K/204R in different countries (Figures 2C and S1B) and observed a consistent IF track (compared to Figure 2B; correlation = 0.96, P value = 1.41e-08). As of March 2021, 203K/204R showed an IF higher than 80% in Europe and Latin America (Figure 2D). In intra-host single nucleotide variation (iSNVs) analyses, we again observed a significant increase in the IF of R203K/G204R (Figure S1C).

Competition and cooperation of 203K/204R and other mutants

The changes in the 203K/204R IF have an increase (I1) from January 2020 to July 2020, a decrease (D) from July 2020 to November 2020, and an increase for the second time (I2) from November 2020 to date (Figure 2B). In the first half of 2020, there was a rapid increase in the IF of both 614G (LG_1) and 203K/204R (LG_3) (Figure 2B). Correlation analyses of IF tracks between pairs of mutants in these three time intervals show that the evolution of R203K/G204R is independent of that of D614G (Figures S2A–S2D; for details, see STAR Methods). The decrease of R203K/G204R in D and the increase in I2 are associated with the increase of A222V in D (correlation = −0.99, P value = 0.0004) and the increase of N501Y in I2 (correlation = 0.99, P value = 0.00016; Table S2A, Figures S2E–S2L, "Cor_-mut.pdf" in Data S1; for details, see STAR Methods), respectively. We tracked the IFs of the combinations ($2^3 = 8$) of the three sets of two-allele polymorphisms (A222V, N501Y, and R203K/G204R) and identified four dominant lineages (Figure 3A). The IFs of the lineages (Figures 3A–3F and S3) indicate that the order of adaptation from high to low is as follows: A222 + 501Y + 203K/204R (AYK), 222V + N501 + R203/G204 (VNR), A222 + N501 + 203K/204R (ANK), and A222 + N501 + R203/G204 (ANR) (Table S2B). The order was confirmed by comparing the growth rates of linages separately in multiple geographically restricted contexts (Figure 3G, Tables S2C and S2D; for details, see STAR Methods).

The lineages showing rapid increases (Figures S4A–S4H), including B.1.1.7, P.1, P.2 (Zeta), P.3 (Theta), and C.37 (Lambda), are all carrying R203K/G204R mutations (Tables S3A and S3B). The overlap of B.1.1.7 and AYK (216,578 strains) accounts for 95.4% of the two lineages' sum-up (226,992 strains). Thus, AYK may be considered as a simplified version of B.1.1.7. For the further elucidation of the relationship between 203K/204R and B.1.1.7, we constructed a phylogenetic tree using all known SARS-CoV-2 strains ("GList_2103.xls" in Data S1). The distribution of lineages along the tree shows that the origin of B.1.1.7 (AYK) was ANK (Figures 3H, S4I, and S4J). It is confirmed by TCS (the method of Templeton, Crandall, and Sing) (Clement et al., 2002) network of all strains (Figure S4K) and further approved by the animations of SARS-CoV-2 evolution lineages (the folder "Animations of trees" in Data S1; the legends follow Figure 3H). The phylogenetic tree also shows that 203R/204K arose independently within more than two sub-lineages and that AYK arose independently within multiple ANK lineages (Figure 3H).

Selection signatures of R203K/G204R

The increase in the IF of 203K/204R suggests adaptiveness (Figure 3G). To evaluate this hypothesis, we performed sliding

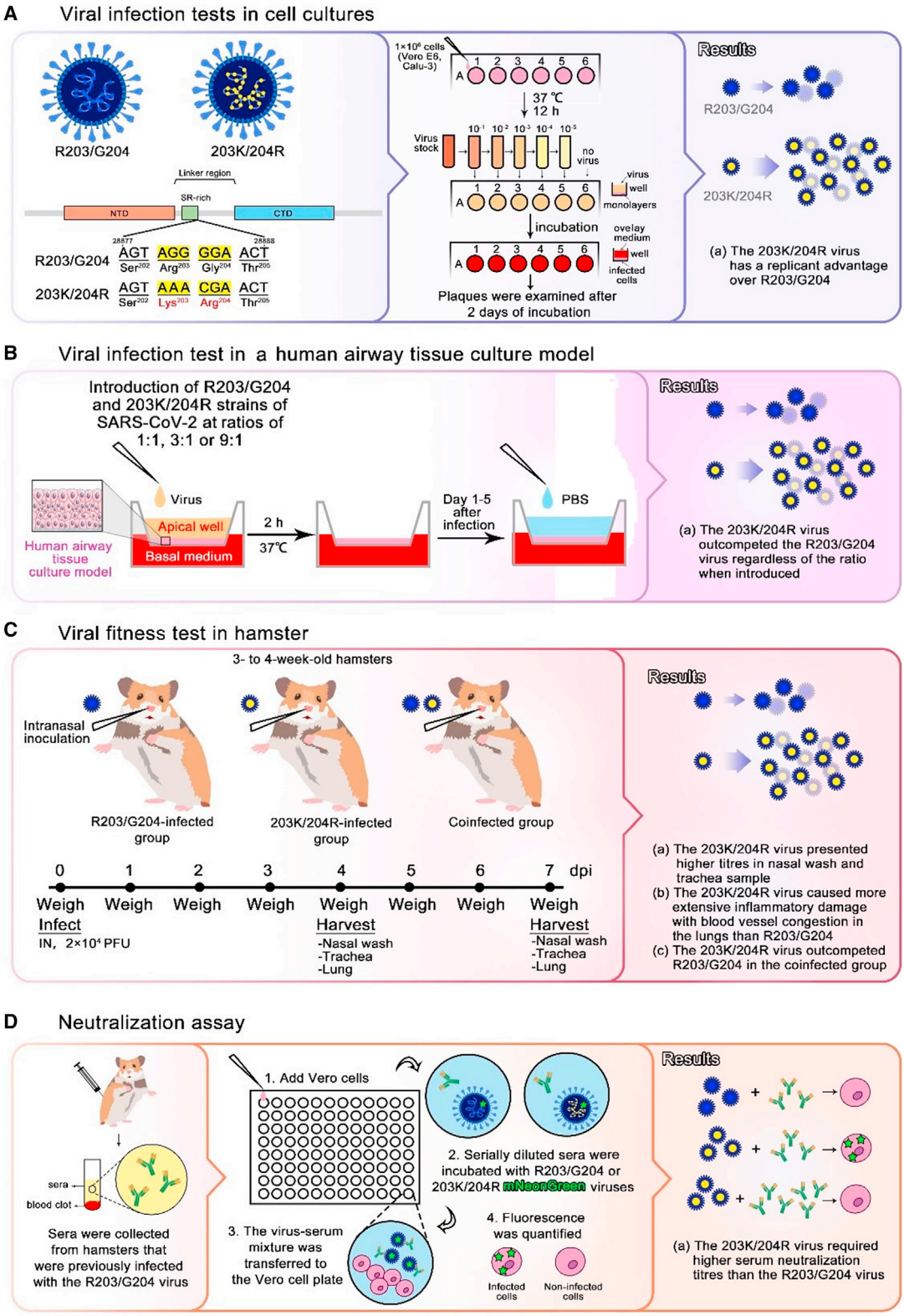


Figure 1. Graphical overview of virology experiments to evaluate the impact of the R203K/G204R mutations on the infectivity and fitness of SARS-CoV-2

The study showed that the R203K/G204R substitutions result in increased virus infectivity and fitness in lung epithelial cells (A), a human airway tissue culture model (B), and the hamster upper airway (C). In addition, 203K/204R showed higher susceptibility when analyzed against serum samples from R203/G204 virus-infected hamsters (D).

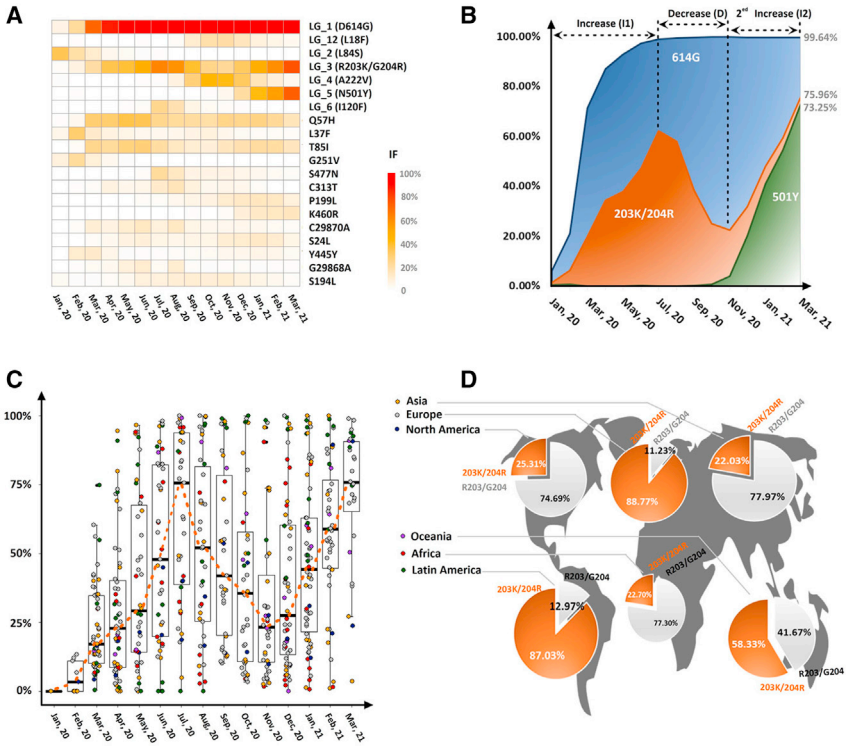


Figure 2. Rapid spread of R203K/G204R

(A) Heatmap showing global IF changes represented by a continuous color gradient from white (0%) to red (100%). Y represents LGs/mutations with an IF > 10% in one month. (B) Global IF tracks of the top three mutants (614G, 510Y and 203K/204R) with the highest IFs recorded to date. The time intervals in which the IFs of 203K/204R increased (I1), decreased (D), and increased for the second time (I2) are annotated at the top. (C) IF changes in different countries over months. The continents of the countries are differentiated by color. An orange dotted line connecting the medians by month was drawn to make the trends discernible. More detailed information is shown in Figure S3 and in the folder "IF in Countries" in Data S1. (D) IFs of 203K/204R (orange) and R203K/G204 (gray) among continents.

window calculations of the composite likelihood ratio (CLR) (Nielsen et al., 2005; Zhu and Bustamante, 2005) in SARS-CoV-2 genomes. A CLR peak surpassing the threshold (the top 5% CLRs in ranking) was considered a positive selection signature with statistical significance. For the convenience of manipulation, we calculated the ratio (m/t) of the CLR at the mutation site (m) and the CLR threshold (t, 0.05) in the genome, where a $CLR_{m/t}$ equal to or higher than 1 indicated a significant CLR peak or adaptive selection signature. We calculated the $CLR_{m/t}$ values for R203/G204 variants and 203K/204R variants. Considering that pattern changes during diversification (Castel et al., 2014) and that the spread of R203K/G204R is not continuous, we calculated the $CLR_{m/t}$ per month. We identified positive selection signatures of the specific mutations. Taking the results from UK strains as an example, there were more CLR peaks (chi-square test P value = 0.098) for 203K/204R variants than R203/G204 variants (Figures 4A and 4B). In July, the increase of 203K/204R reached a peak. In the months near this time point, we observed significantly higher $CLR_{m/t}$ among countries for 203K/204R than for R203/G204 (Figures S5A and S5E). In the comparison of whole-genome diversification between R203/G204 and 203K/204R variants, we found that Tajima's D, Pi, and Theta were lower in 203K/204R variants than in R203/G204 variants (Figures S5B–S5D and S5F).

For a further evaluation of the hypothesis that 203K/204R confers increased transmission fitness, we identified phylogenetic clusters of samples collected in the United Kingdom in time period I1. Clusters of 203K/204R were first detected later than R203/G204 clusters (Figures S5G and S5H, P value = 1.379e-10). We did not find more 203K/204R clusters (306) than R203/G204 clusters (480). The 203K/204R clusters were 21% larger than the R203/G204 clusters on average, but the difference was not significant (P value = 0.1703). Accordingly, the frequency of the sampling of 203K/204R variants continuously increased in I1 (Figure S5I). The UK clusters that were first detected early were larger than those detected later (Figure S5J). These R203K/G204R signatures are similar to previous findings for D614G (Volz et al., 2021).

We further performed a coalescent simulation to evaluate whether the increased frequency of 203K/204R reflected a selective advantage. Following previously applied approaches (Volz et al., 2021), we focused on the sequences of the clusters first detected in January or February and those detected before the end of March, when there was a national lockdown in the United Kingdom; these are considered to represent cocirculating sequences or clusters within the UK (Volz et al., 2021). Simulation with a logistic model showed that the growth rates of cocirculating 203K/204R variants were higher than those of cocirculating R203/G204 ones (P value < 2.2e-16; Figures 4C and S5K). If 203K/204R-infected cases grow exponentially at a rate of r and R203/G204-infected cases grow exponentially at a rate of r(1+s), with s as the estimated mutational selection coefficient, the calculated value of s is 1.82. Simulation with the sky-growth coalescent model for cocirculating sequences or all clusters also showed a higher growth rate of 203K/204R-infected samples than R203/G204-infected samples (Figures S5L and S5M). We did not observe a significant difference between the median growth rates of the 203K/204R and R203/G204 clusters (Figures S5N–S5Q). To evaluate the founder effect, we followed a reported curated susceptible-exposed-infectious-recovered (SEIR) model depicting the relationship between local and other sequences (Volz et al., 2021). We tested the model using 200 sequences sampled randomly from Wales and 100 sequences sampled randomly from outside of Wales in the UK. Using this model, we performed simulations and compared the growth rates between 203K/204R- and R203/G204-infected samples with and without phylogenetic information. The results showed a significantly higher value for 203K/204R than for R203/G204 (Figures 4D and S5R).

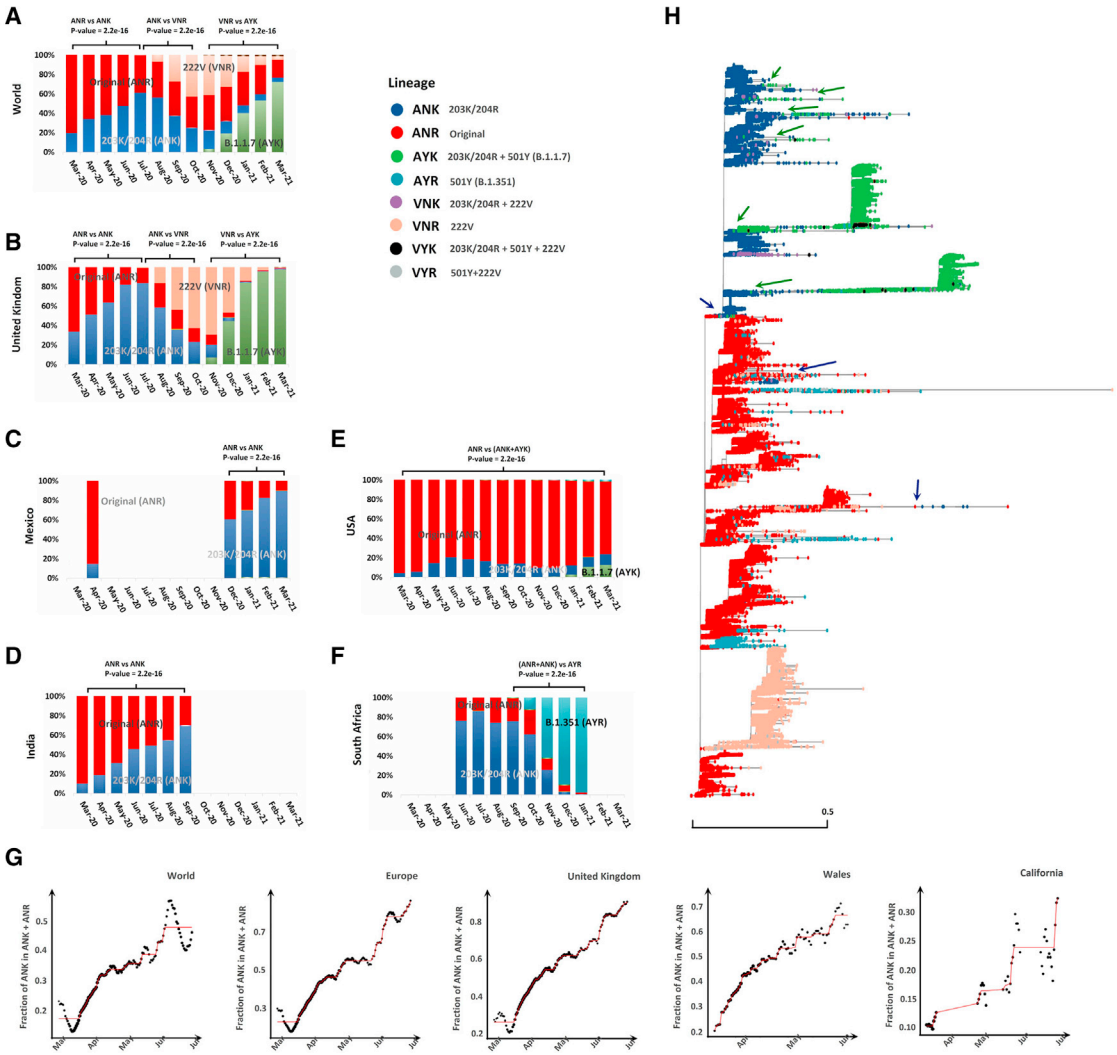


Figure 3. IF changes and the phylogenetic distribution of 8 lineages

Changes in the IF of 8 lineages including A222V (LG_4), N510Y (LG_5), and R203K/G204R (LG_3) and the distribution of lineages in the phylogenetic tree of all strains (H). A to F are the percentage (IF) accumulated area maps of lineages worldwide and in five countries. In B to F, the IFs determined from an insufficient sample size (< 200) are not shown. Lineages are differentiated by color. The legend describing the lineages is in the center. The third column of the legend highlights the mutations carried by each lineage. "Original" denotes the coexistence of the original allele (A222, N501, and R203/G204). In (A) and (B), Fisher's exact tests were performed on the fractions of three pairs of lineages in the beginning and end months of the three time intervals, respectively (ANR versus ANK in I1, ANK versus VNR in D, and VNR versus AYR in I2). In (C) to (E), similar statistics were performed on the fractions of ANR and ANK in the beginning and the end of the months bracketed. In (E), Fisher's exact test was performed for ANR+ANK versus AYR in the beginning and the end of the months bracketed. In (G), there are exemplary sub-figures showing the fitted trend of the fraction change of ANK in ANR+ANK in multiple geographic levels. The time interval is I1. The red folded line is the maximum likelihood estimate with a non-decreasing constraint. The dot size represents the number of sequences on that day. In (H), arrows point to the possible positions of the arising of ANK (blue arrows) and AYK (green arrows).

203K/204R mutant virus show higher fitness than R203/G204 virus

We constructed 203K/204R mutant virus and tested virus replication in hamsters using R203/G204 virus as a control. Hamsters were infected with a 1:1 mixture of R203/G204 and 203K/204R viruses and then assessed for the quantity of the released virus at 4 days and 7 days after infection in a competition experiment. The relative amounts of R203/G204 and 203K/204R viruses were quantified by RT-PCR and Sanger sequencing. The correlation between input PFU (plaque-forming unit) ratios and output RT-PCR amplicon ratios and verification of the actual ratios of R203/G204: 203K/204R achieved upon viral mixing are shown in Figures S6A and S6B. As a result, we observed a higher 203K/204R to R203/G204 ratio, indicating a replication advantage of 203K/204R in hamsters (Figures 4E and 4F). The ratio of 203K/204R to R203/G204 was greater than 1 at 4 and 7 days post infection (dpi), indicating that the 203K/204R virus had a consistent replication advantage over the R203/G204 virus (Figures 4E and 4F).

Moreover, we performed competition experiments in a human airway tissue culture model. After infecting the tissues with a 1:1 ratio of the R203/G204 and 203K/204R viruses, the 203K/204R

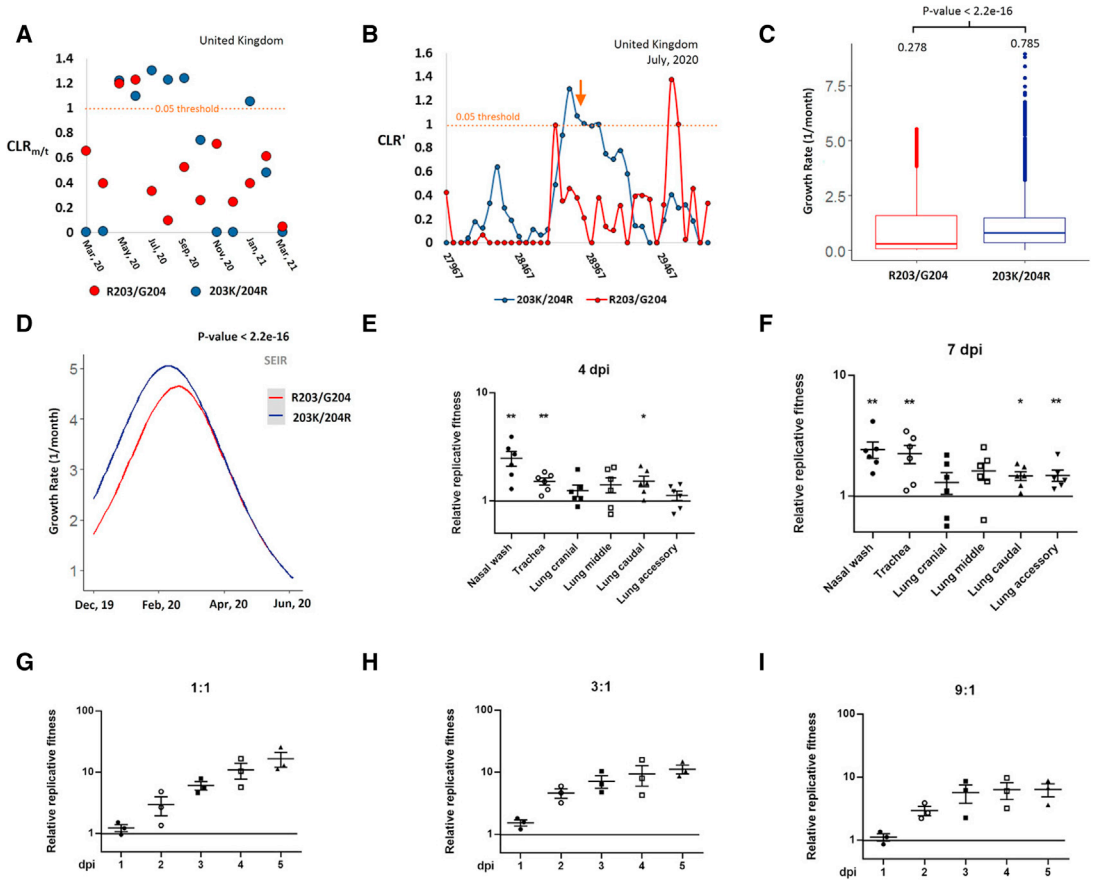


Figure 4. Evidence showing the adaptation of R203K/G204R

(A) Comparison of the CLR $_{m/t}$ per month in the UK between R203/G204 (red) and 203K/204R variants (dark blue). An orange dotted line denotes the top 5% CLR cutoff. (B) A sliding window view showing a CLR peak exceeding the threshold (orange dotted line). This plot is a detailed view of the positive CLR $_{m/t}$ in July in the UK, as shown in (A). CLR' is a transformed CLR value (CLR' = CLR/t). The transformation is to diminish the background difference effects. (C) Comparison of the growth rates simulated in the logistic growth model between R203/G204 (red) and 203K/204R viruses (blue). (D) Fitted curves of the growth rates from the phylodynamic susceptible-exposed-infectious-recovered (SEIR) model. Only genetic data are used. (E and F) Hamsters were inoculated with a 1:1 mixture of the R203/G204 and 203K/204R viruses (104 PFU each). Nasal washes, trachea and lungs were collected on days 4 (E) and 7 (F) after infection. The relative amounts of R203/G204 and 203K/204R RNA were assessed by RT-PCR and Sanger sequencing. Log10 scale is used for the y axis. Data are presented as the mean ± SEM. Dots represent individual hamsters (n = 6). (E) and (F) show that the 203K/204R virus is dominant to the R203/G204 virus in hamsters. (G–I) Competition assay. Mixtures of R203/G204 and 203K/204R viruses with initial ratios of 1:1 (G), 3:1 (H), or 9:1 (I) were inoculated into human airway tissue cultures at a total MOI of 5. Virus ratios after competition were measured by RT-PCR and Sanger sequencing. Log10 scale is used for the y axis. All data are represented as the mean ± SEM. Dots represent individual hamsters (n = 6). *p < 0.05, **p < 0.01. Abbreviation: n.s., nonsignificant.

to R203/G204 ratio increased from 1 dpi to 5 dpi (Figure 4G). In addition, after infecting the airway cultures with a 3:1 or 9:1 ratio of the R203/G204 and 203K/204R viruses, the 203K/204R virus rapidly overcame its initial deficit and showed an advantage over the R203/G204 virus (Figures 4H and 4I). These results demonstrated that the 203K/204R virus could rapidly outcompete the R203/G204 virus in both hamster and human airway tissues, suggesting that the replication advantage of the 203K/204R virus is unrelated to host and tissue specificity.

The 203K/204R virus shows higher infectivity than the R203/G204 virus

The infectivity of the constructed mutant and wild viruses was compared in different cell lines, including the Vero E6 monkey kidney cell line and the Calu-3 human lung epithelial cell line. In Vero E6 cells, the 203K/204R virus replicated with a higher extracellular viral RNA production than R203/G204 at 36 h post infection (hpi) (Figure 5A). The PFU titer and viral subgenomic RNA (sgRNA) were calculated to measure the infectivity of the virus. We found that PFU titer had a similar trend to the extracellular viral RNA levels (Figure 5B), but there were no significant differences in E (envelope protein) sgRNA loads between the two viruses in Vero E6 cells (Figure 5C). Unlike Vero cells, Calu-3 cells infected with the 203K/204R virus produced almost equivalent extracellular viral RNA levels to those detected in cells infected with the R203/G204 virus (Figure 5D). Moreover, the PFU titers and E sgRNA loads of the 203K/204R virus were significantly higher than those of the R203/G204 virus (Figures 5E and 5F), indicating that the R203K/G204R mutation increased the infectivity of SARS-CoV-2 in the human lung cell line.

Thereafter, we performed the same comparison in hamsters. Three- to four-week-old hamsters were infected intranasally with

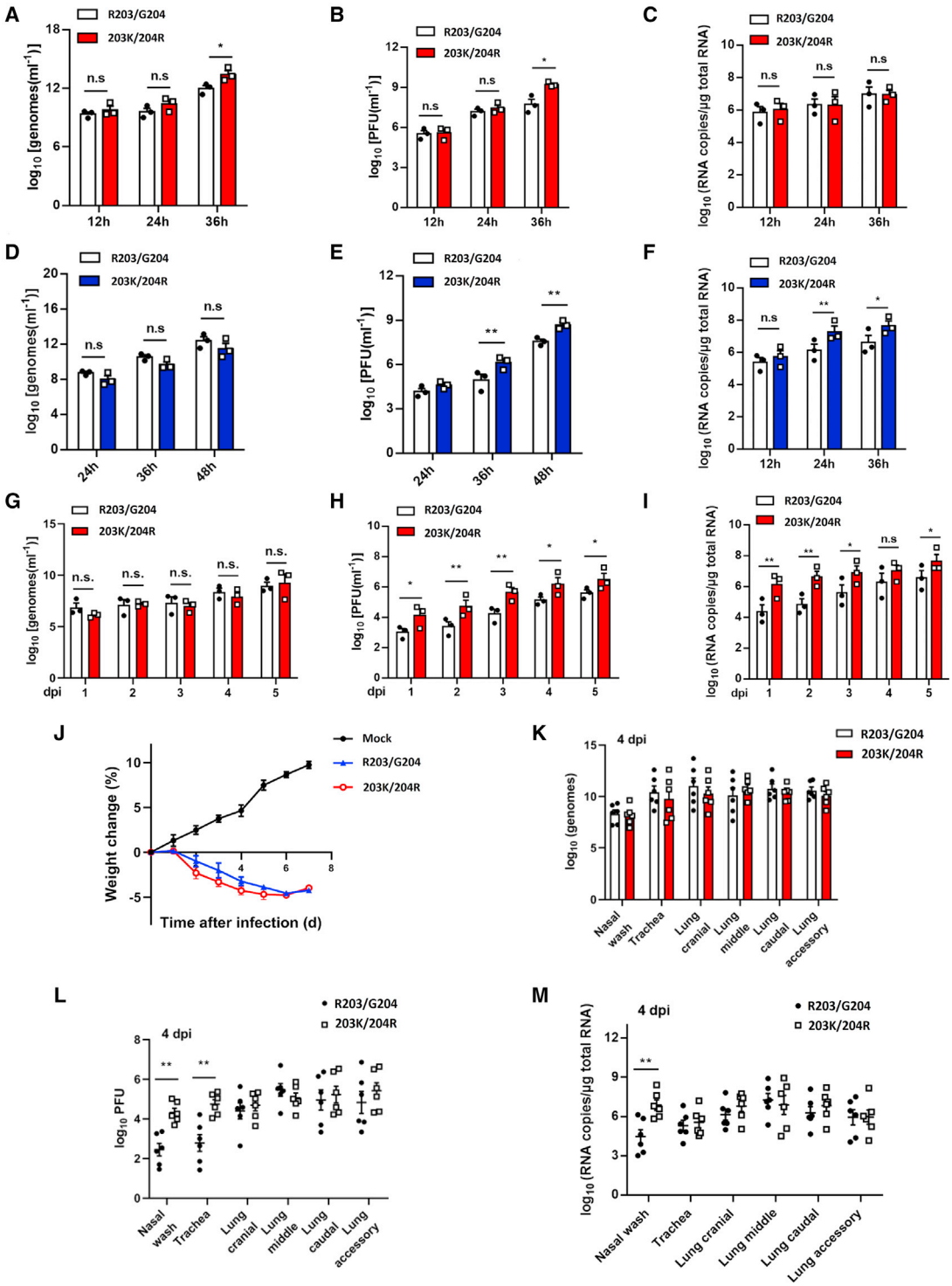


Figure 5. Effects of 203K/204R on viral replication and infectivity

(A–F) Viral replication and viral sgRNA of R203/G204 and 203K/204R viruses produced from Vero E6 (A–C) and Calu-3 (D–F) cell cultures. Cells were infected at a MOI of 0.01. Genomic RNA levels (A) and (D) and infectious viral titers (B) and (E) in the culture medium were determined by plaque assays and qRT-PCR, respectively. The E sgRNA loads (C) and (F) were performed to indicate virion infectivity. Data are represented as mean ± SEM. *$p < 0.05$, **$p < 0.01$. (G–I) Viral replication and viral sgRNA of R203/G204 and 203K/204R viruses. R203/G204 and 203K/204R viruses were inoculated into primary human airway tissues at an MOI of 5. After incubation for 2 h, the cultures were washed with PBS and maintained for 5 days. Genomic RNA levels (G) and infectious viral titers (H) in the culture medium were determined by plaque assays and qRT-PCR, respectively. The E sgRNA loads (I) were calculated to indicate virion infectivity. (G–I) show that the R203K/G204R substitutions enhance SARS-CoV-2 replication in primary human airway tissues. (J–M) Infectivity of R203/G204 and 203K/204R viruses produced

(legend continued on next page)

2×10^4 PFU of the R203/G204 or 203K/204R virus. The infected hamsters from the two groups exhibited similar weight loss (Figure 5J). At 4 dpi, the two viruses produced nearly identical levels of viral RNA across all organs (Figure 5K). We further compared the infectivity of the R203/G204 and 203K/204R viruses produced in hamsters by determining their PFU titers levels and viral sgRNA levels. Similar to the previous observation, the infectious viral titers measured in nasal wash and trachea samples, as well as the E sgRNA loads in nasal wash samples from hamsters infected with the 203K/204R virus, were significantly higher than those in hamsters infected with the R203/G204 virus at 4 dpi (Figures 5L and 5M). Another independent repeat of hamster experiments infected with R203/G204 and 203K/204R viruses confirmed the above results (Figures S6C–S6F). We did not observe significant difference at 7 dpi (Figures S6G–S6I).

In comparisons conducted in a human airway model, there were no differences in viral RNA yields between the variants (Figure 5G). However, the PFU titers and viral sgRNA levels of 203K/204R virus were significantly higher than those of R203/G204 virus (Figures 5H and 5I). These results, combined together, demonstrated that the R203K/G204R mutations enhanced viral replication efficiency and further increased the virion infectivity commonly. Consistent with the experimental results, we previously observed a negative correlation of the R203/G204 IF with the cycle threshold for a positive signal in E gene-based RT-PCR (Ct) (Zhu et al., 2021b).

In order to test the sensitivity of the mutant virus to the neutralization serum, we measured the neutralization titers of a panel of sera collected from hamsters that were infected with the R203/G204 virus. Each serum sample was analyzed using mNeonGreen reporter R203/G204 or 203K/204R viruses. All sera exhibited 1.14- to 2.08-fold higher neutralization titers (mean 1.51-fold) against the heterologous 203K/204R virus than against the homologous R203/G204 virus (Figures S6J–S6R). Serum 3 presented the highest neutralization titer (Figure S6L). The results suggested that the R203K/G204R mutant virus confers higher susceptibility to serum neutralization.

The 203K/204R virus shows an association with increased disease severity

In experiments, we observed that hamsters infected with the 203K/204R virus showed more extensive inflammatory damage with blood vessel congestion in the lungs than hamsters infected with R203/G204 (Figures 6A–6C). For further validation, the statistical analyses of sequenced strains with patient information were performed (Tables S4A and S4B, from GISAID). We found that the 203K/204R virus showed significant increases in the symptomatic:asymptomatic, hospitalized:outpatient, mild:severe, and deceased:released ratios (Figures 6D–6G). To avoid bias resulting from geographical differences, we calculated the statistics of data collected on a small scale (Figures 6D–6G and S6T–S6W). Through correlation analyses between the 203K/204R IF and the case fatality rate (CFR) per month in different countries (Table S4C), we found a significant positive correlation between the 203K/204R IF and CFR (Figures 6H–6J). These statistical analyses confirmed the increased clinical virulence of 203K/204R mutant virus.

Structural implications of the R203K/G204R mutations for RNP assembly

To investigate the impact of R203K/G204R on the tertiary structure of the nucleocapsid, we built a structural model of the N protein based on documented cryo-EM reports (Gui et al., 2017; Figure 7). We found that the N-terminal domain (NTD) and C-terminal domain (CTD) were docked in a reverse L-shape manner in the murine hepatitis virus (MHV) nucleocapsid electron density map (Gui et al., 2017; Figure 7A). According to a previous study (Chen et al., 2007), the CTD dimer is suggested to be an assembly unit of N. The R203K/G204R mutations are located in the linker region of N. The linker region is a serine/arginine-rich region with high flexibility. Due to the abundance of arginine, the linker region is more alkaline than the NTD and CTD (Figure 7B)—i.e., the pI (isoelectric point) of NTD and CTD is around 10, whereas the pI of the linker region reaches 11.88. As lysine is also a basic amino acid, the R203K mutation causes only a small change in the pI of the linker region. In contrast, G204R introduces an additional basic amino acid, leading to a more significant increase in the linker region pI. The electrostatic surface potential of N shows that the CTD dimer, also known as the RNA-binding dimerization domain, is rich in positive charges (Figure 7C). However, owing to the lack of available 3D structures, the existing surface electrostatic potential analyses have ignored the linker region. The linker region is highly basic (i.e., positively charged), which suggests that this region is also involved in the RNA binding process. Thus, the R203K/G204R mutations may impact virus RNP assembly.

A more complete nucleocapsid model was also proposed based on cryo-EM reports (Gui et al., 2017). Four N dimers are packed head to head, forming an "X" shape (Figure 7D), and these structures are then packaged into a helical filament (Figure 7E). The CTD dimer forms the core of the helical nucleocapsid, and the NTDs form the two arms outside of the building block. Based on the surface electrostatic potential of the proposed model, the possible RNA binding groove of the helical nucleocapsid model was identified (Figure 7E). According to the winding path of the RNA, the RNA also surrounds and possibly interacts with the linker region between the core (CTD) and the arms (NTD). Combined with the abovementioned findings that the R203K/G204R mutations could impact the local charge of the N protein, the R203K/G204R mutations may promote the binding of RNA by increasing the positive charge within the linker region to increase the RNP assembly efficiency, thereby accelerating the virus' replication.

DISCUSSION

In our previous work, we observed the rapid spread of R203K/G204R mutation in the initial four months after the onset of

in hamsters. Three- to four-week-old hamsters were infected intranasally with 2×10^4 PFU of the R203/G204 or 203K/204R virus or PBS (mock). All data came from a single experiment. Weight loss (J) was monitored for 7 dpi. Data are presented as mean ± SEM; n = 12 (all cohorts) at days 0–4; n = 6 (all cohorts) at days 5–7. Weight loss was analyzed by two-factor analysis of variance (ANOVA) with Tukey's post hoc test. Amounts of viral genomes (K) and infectious titers (L) were quantified in nasal wash, trachea, and lung samples on the 4th dpi. Dots represent individual hamsters (n = 6). The E sgRNA loads (M) at 4 dpi were calculated as a measurement of infectivity. Dots represent individual hamsters (n = 6). Data are presented as the mean ± SEM. **p < 0.01.

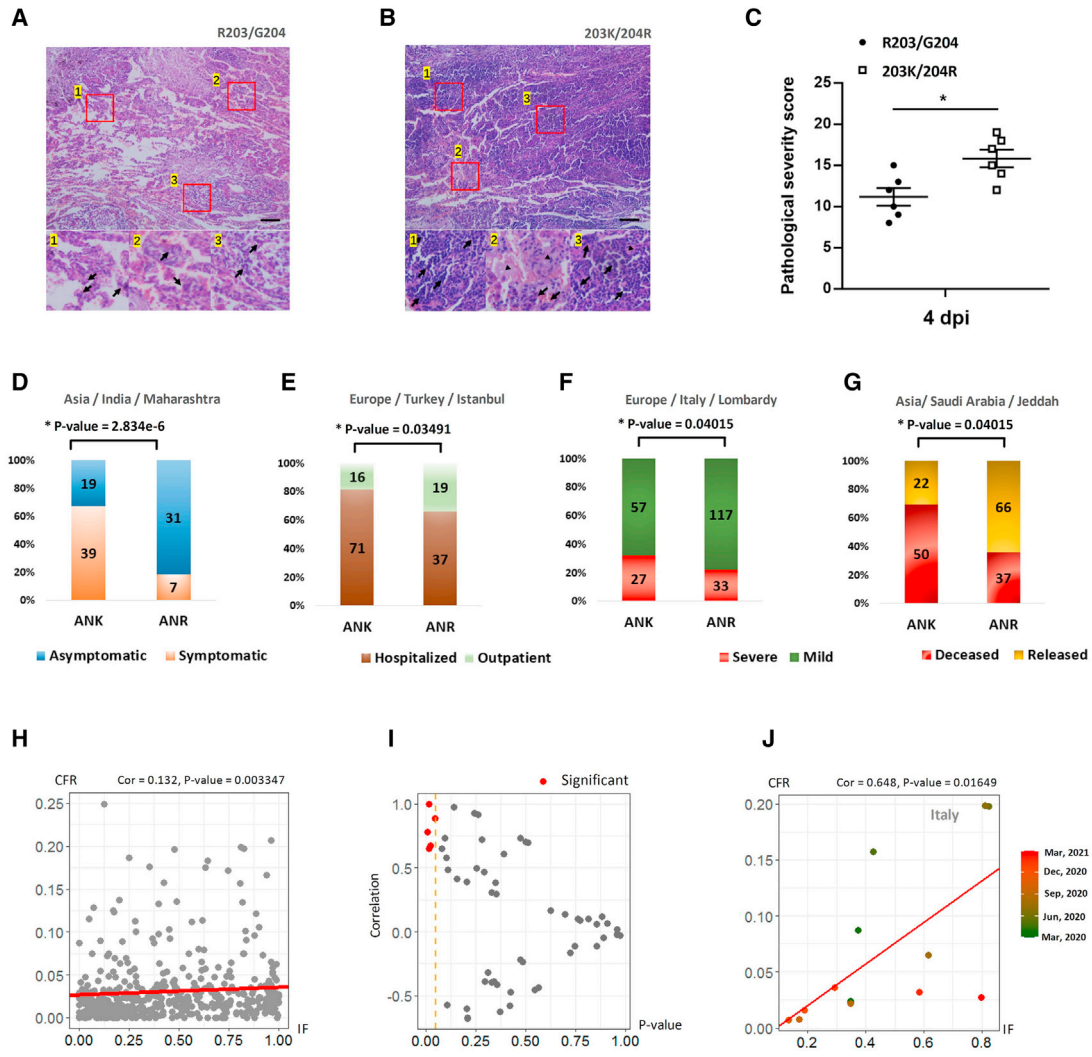
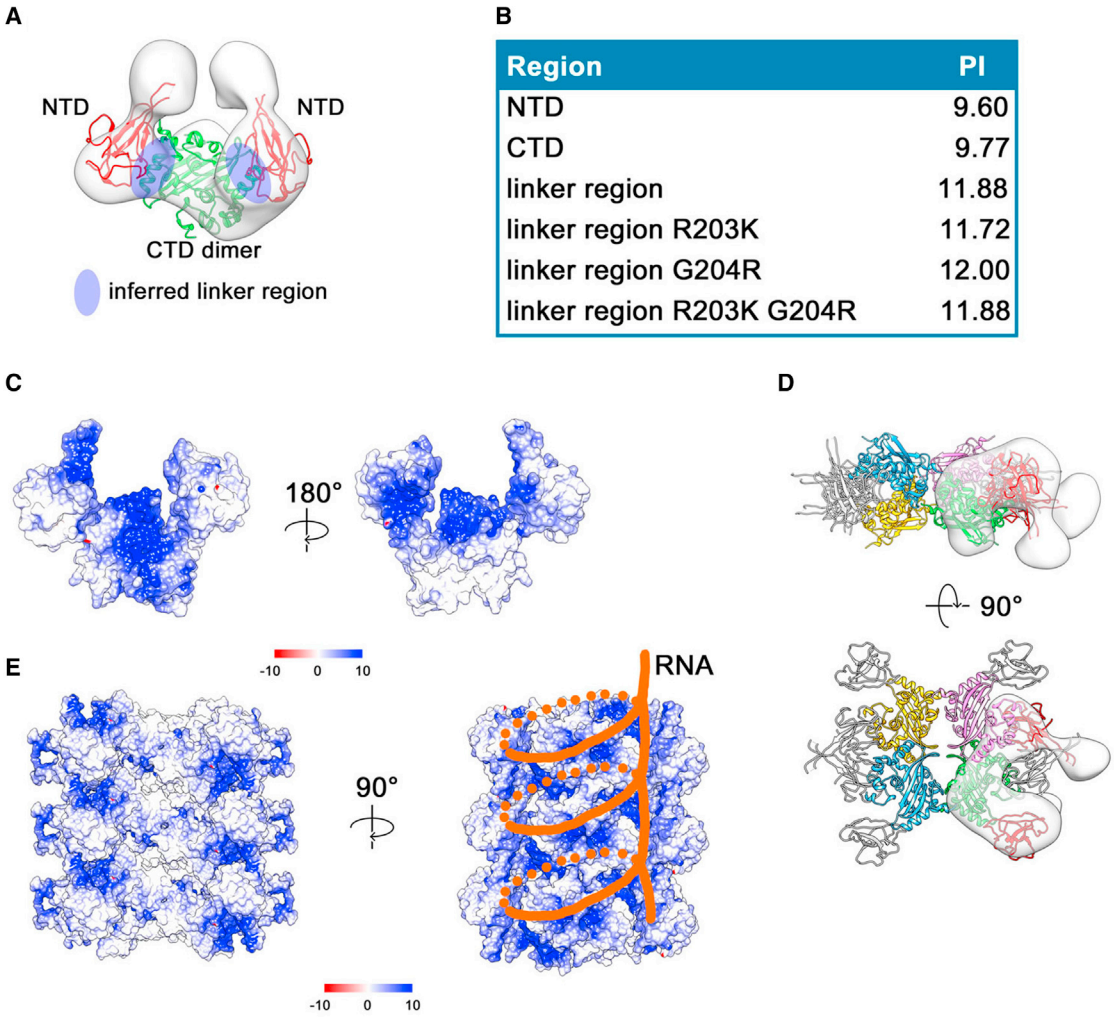


Figure 6. Effects of R203K/G204R on the severity of disease

(A and B) Severity of infections caused by R203/G204 and 203K/204R. Haemotoxylin and eosin (H&E) staining of lung sections from R203/G204 and 203K/204R virus-infected hamsters collected at 4 dpi. The lower photographs are higher magnification images of the regions denoted by rectangles in the upper photographs. The upper panel shows inflammatory damage with blood vessel congestion. The lower panel shows bronchioles with aggregation of inflammatory cells (arrow) and surrounding alveolar wall infiltration (arrowhead). Scale bar, 100 μm. (C) Histopathology score of lung sections. Lung lobes were scored individually using the following scoring system: 0, no pathological change; 1, affected area (£10%); 2, affected area (> 10% and £30%); 3, affected area (> 30% and £50%); and 4, affected area (> 50%). We obtained 5 slices from each hamster and the scores of lung slices were added to determine the total pathology score per animal. Dots represent individual hamsters (n = 6). (D–G) Prediction of the clinical outcomes of ANR (R203/G204) and ANK (203K/204R). Four pairs of colors denote four pairs of opposite patient statuses. Collection sites are listed at the top of the figures. The y axis shows the ratios between lineages with opposite patient statuses. Lineage numbers are provided. We tested the significance of the change in the ratio by the chi-square test. (H–J) Correlation analysis results between mutant IFs and CFRs. (H) Correlation analysis between the ratios of mutants (x axis) and the median CFRs of COVID-19 among months and in different countries (y axis). The p value was calculated using Spearman's test. (I) Scatter map showing the distribution of the IF-CFR correlations (y axis) and the p values (x axis) in different countries. A vertical orange dotted line denotes the 0.05 p value cutoff. (J) A specific example showing the correlation of mutant IFs and CFRs in a single country, Italy. Continuously changing colors denote the months of the IF-CFR pairs.

the SARS-CoV-2 pandemic and predicted that these mutations may benefit virus replication based on statistics (Zhu et al., 2021b). In this study, we have performed comprehensive statistical analyses of R203K/G204R and confirmed the adaptiveness of the mutations. We further experimentally verified the increase in infectivity of the mutant virus and proved the increase in virulence and fitness of SARS-CoV-2 virus introduced by the N protein mutations (Figure 1). In comparison with the original R203/G204 virus, the 203K/204R virus exhibited increased viral replication in competition assay not only in different cell lines but also in primary human upper airway tissues and hamsters. We found that the increase in infectivity was due to the promotion of virus replication efficiency, which may have been caused by the change in the local charge of the N protein resulting from the R203K/G204R mutations. The G203K/G204R mutant may increase the assembly efficiency



Region	PI
NTD	9.60
CTD	9.77
linker region	11.88
linker region R203K	11.72
linker region G204R	12.00
linker region R203K G204R	11.88

Figure 7. Tentative model of the SARS-CoV-2 nucleocapsid

(A) Docking of the crystal structures of the SARS-CoV-2 nucleocapsid protein NTD monomer (PDB ID: 6M3M) and CTD dimer (PDB ID: 6WZQ) into the 3D density map of the MHV nucleocapsid. The NTD is colored red. The CTD dimer is colored green. The inferred linker region is highlighted in blue. The theoretical pI of each region is shown in (B). The electrostatic surface potential corresponding to (A) is shown in (C). (D) Four nucleocapsid protein dimer packages in an asymmetrical unit of the nucleocapsid. CTD dimers are colored blue, yellow, magenta, and green, respectively. A density map of one asymmetry subunit is shown as a gray, semitransparent surface with the corresponding fitted ribbon models colored green (CTDs) and red (NTDs). The NTDs of the other three nucleocapsid protein dimers are colored gray. Based on the model proposed in (C) and (D), we built a tentative model of the nucleocapsid (E). In (D) and (E), the positively charged region on the surface is colored blue, as shown in the scale bar, −10 to 10 kcal/(mol·e-). The possible RNA binding grooves are indicated by orange lines.

of RNP and further increase the assembly efficiency of the virus particle, thus increasing the quality of the virus particle (PFU) instead of the quantity (genomic RNA level) of the virus particle (Figure 5). The increase in replication efficiency may have ultimately led to the increases in virulence and fitness. Other high-infectivity-related mutations, such as D614G and N510Y, also show associations with increases in fitness and disease severity (Biswas and Mudi, 2020; Liu et al., 2021; Plante et al., 2021; Zhao et al., 2021b; Zhu et al., 2021b). Like D614G, R203K/G204R is associated with predominance (Figure 3) and is shared by the rapidly increasing lineages B.1.1.7 and P.1. Understanding the effects of these mutations will likely be important for the prevention of further SARS-CoV-2 infections.

Since the onset of the SARS-CoV-2 pandemic, more attention has been focused on S mutations than mutations in other virus components. Our findings indicated that N mutations also alter the function and fitness of viruses. Although the S protein and N protein, both structural proteins, are located in different parts of a virion, there are many similarities between D614G and R203K/G204R related to the consequences of the virus properties. R203K/G204R variants also show increased infectivity and fitness and an association with the severity of disease. The mutant 614G is reported to result in a higher neutralization titers than D614 (Plante et al., 2021). Additionally, the 203K/204R virus is more susceptible to neutralization than the R203/G204 virus. In principle, the increased sensitivity of virus to neutralizing antibody may be the result of decreased assembled S proteins on

the virion surface because of a more efficient assembly of virus. Among the 96 mutations we identified, 15 (15.6%) were N mutations and 14 (14.6%) were S mutations (Table S1A). In B.1.1.7, one quarter (7/28) of the mutations were N mutations and one quarter (7/28) were S mutations (Table S3B). Hence, N mutations should be considered equally important in future efforts.

Although 203K/204R variants show higher fitness than the preceding R203/G204 variants, the IFs of 203K/204R did not increase continuously. This is not similar to findings for 614G. The reason may be the cooperation and competition between 203K/204R and other mutants, or it may be the result of the increased sensitivity to neutralizing antibody for 203K/204R mutants. We did not observe positive signatures of 203K/204R in some months or countries (Figures 4A and S4A–S4D), which may have been due to the effects of related adaptive mutants. Possibly for the same reason, we observed positive selection signatures of R203/G204 variants.

The increases in the infectivity and virulence of R203K/G204R variants could contribute to the increased transmission and mortality of B.1.1.7 (Davies et al., 2021; Washington et al., 2021) and the increased severity of disease associated with P.1 (Funk et al., 2021; Martins et al., 2021). The recently evolved Indian lineages B.1.617.1 (Kappa) (Yadav et al., 2021), B.1.617.2 (Delta), and AY.1 (Delta-plus) also carry a point mutation at 28,881, which is a novel N mutation, R203M, instead of R203K/G204R (Table S3A). This newly emerging mutation at the same location on the N protein may bolster similar functional effects, implying the biological importance of 28,881 in the nucleocapsid region of the SARS-CoV-2 genome. Analyses in IF with the latest SARS-CoV-2 sequence data (from April 2021 to September 2021, updated to 3,438,667 sequences; ''GList_2109.xls'' in Data S1) show that 203M/G204 have become predominant at present (Figures S7A–S7C), possibly due to the high replication efficiency and strong immune evasion capability of the lineage Delta (Mlcochova et al., 2021). Like the high correlation between 203K/204R and B.1.1.7, 203M/G204 has a high correlation with Delta (Figure S7D).

In summary, we identified the adaptation of the N protein mutations R203K/G204R through thorough *in silico* evolutionary analyses. Using the authentic virus, we proved that R203K/G204R increase viral replication, which enhances the infectivity, fitness, and virulence of SARS-CoV-2. R203K/G204R increase the sensitivity of virus to neutralizing antibodies, which may be complemented by immune resistance mutations such as N501Y and E484K. These findings provide important information to understand the fast-evolving SARS-CoV-2 virus and contribute to the control of the pandemic. However, further studies are needed to address the functional nucleocapsid mutations, e.g., the combinational effects of the nucleocapsid mutations and the mutations in the spike protein.

LIMITATIONS OF THE STUDY

We have performed analyses and experiments on R203K/G204R mutant virus. However, the detailed mechanism of these mutations is still ambiguous, and it may require more extensive biochemical and structural research in the future. Selection signature for R203K/G204R did not perform consistently across the whole period of the pandemic like D614G. Further surveying and monitoring of these mutations in the future may be needed.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Animal husbandry and infection
  - Cell culture and infection
- METHOD DETAILS
  - Identification and statistics of SARS-CoV-2 mutations
  - Analyses of the confounding effects of 203K/204R and other mutants
  - Evolutionary analysis
  - Function prediction based on clinical and epidemic data
  - Generation of the R203K/G204R mutant virus
  - Plaque assays
  - Neutralization assay
  - Viral infection in a primary human airway tissue model
  - Validation of competition assay
  - Viral sgRNA assay and genomic RNA assay
  - Pathological examination
  - 3D structural model of assembled nucleocapsid
  - Haplotype network
- QUANTITATIVE AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2021.11.005.

ACKNOWLEDGMENTS

We gratefully acknowledge the submitting and the originating laboratories where genetic sequence data were generated and shared via NCBI and the GISAID Initiative. This work was supported by grants from the National Natural Science Foundation of China, SGC's Rapid Response Funding for COVID-19 (C-0002), the National Key Research and Development Program (2019YFC1604600), the National Natural Science Foundation of China (81970008, 32170661, and 82000020), the Fundamental Research Funds for the Central Universities (2021CDJYGRH-009), the Youth Innovative Talents Training Project of Chongqing (CY210102), and the National Natural Science Foundation of HeBei province (19226631D).

AUTHOR CONTRIBUTIONS

Z.Z., G.M., H.W., K.M., P.D., W.T., Y.X., G.L., and H.L. collected the data, performed population genetic analyses, and took part in the editing of the manuscript. N.X., H.W., B.F., W.X., and X.L. performed the experiments. G.M., K.M., Z.Z., and W.Z. performed the protein structure analysis. Z.Z. and G.M. conceived the idea. Z.Z., G.M., and X.L. wrote the manuscript. Z.Z., X.L., and G.M. coordinated the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 26, 2021
Revised: October 7, 2021
Accepted: November 9, 2021
Published: November 13, 2021

REFERENCES

Balaban, M., Moshiri, N., Mai, U., Jia, X., and Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS ONE *14*, e0221068.

Biswas, S.K., and Mudi, S.R. (2020). Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19. Genomics Inform. *18*, e44.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. *15*, e1006650.

C Caserta, L., Mitchell, P.K., Plocharczyk, E., and Diel, D.G. (2021). Identification of a SARS-CoV-2 Lineage B1.1.7 Virus in New York following Return Travel from the United Kingdom. Microbiol Resour Announc *10*, e00097-e21.

Castel, G., Razzauti, M., Jousselin, E., Kergoat, G.J., and Cosson, J.F. (2014). Changes in diversification patterns and signatures of selection during the evolution of murinae-associated hantaviruses. Viruses *6*, 1112–1134.

Chen, C.-Y., Chang, C.K., Chang, Y.-W., Sue, S.-C., Bai, H.-I., Riang, L., Hsiao, C.-D., and Huang, T.H. (2007). Structure of the SARS coronavirus nucleocapsid protein RNA-binding dimerization domain suggests a mechanism for helical packaging of viral RNA. J. Mol. Biol. *368*, 1075–1086.

Cheng, L., Song, S., Zhou, B., Ge, X., Yu, J., Zhang, M., Ju, B., and Zhang, Z. (2021). Impact of the N501Y substitution of SARS-CoV-2 Spike on neutralizing monoclonal antibodies targeting diverse epitopes. Virol. J. *18*, 87.

Clement, M., Snell, Q., Walker, P., Posada, D., and Crandall, K. (2002). TCS: Estimating gene genealogies. Parallel and Distributed Processing Symposium. International Proceedings *2*, 184.

Collier, D.A., De Marco, A., Ferreira, I.A.T.M., Meng, B., Datir, R.P., Walls, A.C., Kemp, S.A., Bassi, J., Pinto, D., Silacci-Fregni, C., et al.; CITIID-NIHR BioResource COVID-19 Collaboration; COVID-19 Genomics UK (COG-UK) Consortium (2021). Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. Nature *593*, 136–141.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M.L., et al. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. *25*, 2000045.

Dagotto, G., Mercado, N.B., Martinez, D.R., Hou, Y.J., Nkolola, J.P., Carnahan, R.H., Crowe, J.E., Jr., Baric, R.S., and Barouch, D.H. (2021). Comparison of Subgenomic and Total RNA in SARS-CoV-2 Challenged Rhesus Macaques. J. Virol. *95*, e02370-e20.

Davies, N.G., Jarvis, C.I., Edmunds, W.J., Jewell, N.P., Diaz-Ordaz, K., and Keogh, R.H.; CMMID COVID-19 Working Group (2021). Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. Nature *593*, 270–274.

DeGiorgio, M., Huber, C.D., Hubisz, M.J., Hellmann, I., and Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics *32*, 1895–1897.

Dejnirattisai, W., Zhou, D., Supasa, P., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., Tuekprakhon, A., Nutalai, R., et al. (2021). Antibody evasion by the P.1 strain of SARS-CoV-2. Cell *184*, 2939–2954.e9.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Faria, N.R., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D.D.S., Mishra, S., Crispim, M.A.E., Sales, F.C.S., Hawryluk, I., McCrone, J.T., et al. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science *372*, 815–821.

Funk, T., Pharris, A., Spiteri, G., Bundle, N., Melidou, A., Carr, M., Gonzalez, G., Garcia-Leon, A., Crispie, F., O'Connor, L., et al.; COVID study groups (2021). Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. Euro Surveill. *26*, 2100348.

Garcia-Beltran, W.F., Lam, E.C., St Denis, K., Nitido, A.D., Garcia, Z.H., Hauser, B.M., Feldman, J., Pavlovic, M.N., Gregory, D.J., Poznansky, M.C., et al. (2021). Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell *184*, 2372–2383.e9.

Gui, M., Liu, X., Guo, D., Zhang, Z., Yin, C.C., Chen, Y., and Xiang, Y. (2017). Electron microscopy studies of the coronavirus ribonucleoprotein complex. Protein Cell *8*, 219–224.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

Hoffmann, M., Arora, P., Groß, R., Seidel, A., Hörnich, B.F., Hahn, A.S., Krüger, N., Graichen, L., Hofmann-Winkler, H., Kempf, A., et al. (2021). SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. Cell *184*, 2384–2393.e12.

Hou, Y.J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K.H., 3rd, Leist, S.R., Schäfer, A., Nakajima, N., Takahashi, K., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science *370*, 1464–1468.

Hussain, M., Jabeen, N., Raza, F., Shabbir, S., Baig, A.A., Amanullah, A., and Aziz, B. (2020). Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein. J. Med. Virol. *92*, 1580–1586.

Hutter, S., Vilella, A.J., and Rozas, J. (2006). Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics *7*, 409.

Jangra, S., Ye, C., Rathnasinghe, R., Stadlbauer, D., Krammer, F., Simon, V., Martinez-Sobrido, L., García-Sastre, A., and Schotsaert, M.; Personalized Virology Initiative study group (2021). SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. Lancet Microbe *2*, e283–e284.

Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L.E., Brookes, D.H., Wilson, L., Chen, J., Liles, K., et al. (2018). Improvements to the APBS biomolecular solvation software suite. Protein Sci. *27*, 112–128.

Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., and Chang, H. (2020). The Architecture of SARS-CoV-2 Transcriptome. Cell *181*, 914–921.e10.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al.; Sheffield COVID-19 Genomics Group (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell *182*, 812–827.e19.

Langdon, W.B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min. *8*, 1.

Leigh, J.W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. Methods Ecol. Evol. *6*, 1110–1116.

Lewontin, R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics *49*, 49–67.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Liu, Y., Liu, J., Plante, K.S., Plante, J.A., Xie, X., Zhang, X., Ku, Z., An, Z., Scharton, D., Schindewolf, C., et al. (2021). The N501Y spike substitution enhances SARS-CoV-2 transmission. bioRxiv, 2021.03.08.434499.

Martins, A.F., Zavascki, A.P., Wink, P.L., Volpato, F.C.Z., Monteiro, F.L., Rosset, C., De-Paris, F., Ramos, A.K., and Barth, A.L. (2021). Detection of SARS-CoV-2 lineage P.1 in patients from a region with exponentially increasing hospitalisation rate, February 2021, Rio Grande do Sul, Southern Brazil. Euro Surveill. *26*, 2100276.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mercatelli, D., and Giorgi, F.M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol. *11*, 1800.

Mlcochova, P., Kemp, S.A., Dhar, M.S., Papa, G., Meng, B., Ferreira, I.A.T.M., Datir, R., Collier, D.A., Albecka, A., Singh, S., et al.; Indian SARS-CoV-2 Genomics Consortium (INSACOG); Genotype to Phenotype Japan (G2P-Japan) Consortium; CITIID-NIHR BioResource COVID-19 Collaboration (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. Nature *599*, 114–119. https://doi.org/10.1038/s41586-021-03944-y.

Mok, B.W.-Y., Cremin, C.J., Lau, S.-Y., Deng, S., Chen, P., Zhang, A.J., Lee, A.C.-Y., Liu, H., Liu, S., Ng, T.T.-L., et al. (2020). SARS-CoV-2 spike D614G variant exhibits highly efficient replication and transmission in hamsters. bioRxiv, 2020.08.28.271635.

Morton, N.E. (1955). Sequential tests for the detection of linkage. Am. J. Hum. Genet. *7*, 277–318.

Mwenda, M., Saasa, N., Sinyange, N., Busby, G., Chipimo, P.J., Hendry, J., Kapona, O., Yingst, S., Hines, J.Z., Minchella, P., et al. (2021). Detection of B.1.351 SARS-CoV-2 Variant Strain - Zambia, December 2020. MMWR Morb. Mortal. Wkly. Rep. *70*, 280–282.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. Genome Res. *15*, 1566–1575.

Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., Fontes-Garfias, C.R., et al. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. Nature *592*, 116–121. Published online October 26, 2020.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

Ralph, R., Lew, J., Zeng, T., Francis, M., Xue, B., Roux, M., Toloue Ostadgavahi, A., Rubino, S., Dawe, N.J., Al-Ahdal, M.N., et al. (2020). 2019-nCoV (Wuhan virus), a novel Coronavirus: human-to-human transmission, travel-related cases, and vaccine readiness. J. Infect. Dev. Ctries. *14*, 3–17.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. *67*, 901–904.

Rochman, N.D., Wolf, Y.I., Faure, G., Mutz, P., Zhang, F., and Koonin, E.V. (2021). Ongoing Global and Regional Adaptive Evolution of SARS-CoV-2. bioRxiv, 2020.10.12.336644.

Rueca, M., Bartolini, B., Gruber, C.E.M., Piralla, A., Baldanti, F., Giombini, E., Messina, F., Marchioni, L., Ippolito, G., Di Caro, A., and Capobianchi, M.R. (2020). Compartmentalized Replication of SARS-Cov-2 in Upper vs. Lower Respiratory Tract Assessed by Whole Genome Quasispecies Analysis. Microorganisms *8*, 1302.

Salvatori, G., Luberto, L., Maffei, M., Aurisicchio, L., Roscilli, G., Palombo, F., and Marra, E. (2020). SARS-CoV-2 SPIKE PROTEIN: an optimal immunological target for vaccines. J. Transl. Med. *18*, 222.

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. *22*, 30494.

Slatkin, M. (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485.

Smith, E.C., Blanc, H., Surdel, M.C., Vignuzzi, M., and Denison, M.R. (2013). Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog. *9*, e1003565.

Stajich, J.E. (2007). An Introduction to BioPerl. Methods Mol. Biol. *406*, 535–548.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. *4*, vey016.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

Tajima, F. (1993). Measurement of DNA polymorphism. In Mechanisms of Molecular Evolution, N. Takahata and A.G. Clark, eds. (Japan Scientific Societies Press, Tokyo and Sinauer Associates, Inc), pp. 37–59.

Trucchi, E., Gratton, P., Mafessoni, F., Motta, S., Cicconardi, F., Mancia, F., Bertorelle, G., D'Annessa, I., and Di Marino, D. (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.

Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect. Dis. *20*, 669–677.

Vilella, A.J., Blanco-Garcia, A., Hutter, S., and Rozas, J. (2005). VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics *21*, 2791–2793.

Volz, E.M., and Siveroni, I. (2018). Bayesian phylodynamic inference with complex models. PLoS Comput. Biol. *14*, e1006546.

Volz, E., Hill, V., McCrone, J.T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F.F., et al.; COG-UK Consortium (2021). Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell *184*, 64–75.e11.

Wang, Y., Wang, D., Zhang, L., Sun, W., Zhang, Z., Chen, W., Zhu, A., Huang, Y., Xiao, F., Yao, J., et al. (2020). Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2 Population in COVID-19 Patients. bioRxiv, 2020.05.20.103549.

Wang, P., Nair, M.S., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., Yu, J., Zhang, B., Kwong, P.D., et al. (2021). Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature *593*, 130–135.

Washington, N.L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E.T., Schiabor Barrett, K.M., Larsen, B.B., Anderson, C., White, S., Cassens, T., et al. (2021). Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. Cell *184*, 2587–2594.e7.

Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature *581*, 465–469.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Xie, X., Muruato, A., Lokugamage, K.G., Narayanan, K., Zhang, X., Zou, J., Liu, J., Schindewolf, C., Bopp, N.E., Aguilar, P.V., et al. (2020). An Infectious cDNA Clone of SARS-CoV-2. Cell Host Microbe *27*, 841–848.e3.

Xie, X., Lokugamage, K.G., Zhang, X., Vu, M.N., Muruato, A.E., Menachery, V.D., and Shi, P.Y. (2021). Engineering SARS-CoV-2 using a reverse genetic system. Nat. Protoc. *16*, 1761–1784.

Yadav, P.D., Sapkal, G.N., Abraham, P., Ella, R., Deshpande, G., Patil, D.Y., Nyayanit, D.A., Gupta, N., Sahay, R.R., Shete, A.M., et al. (2021). Neutralization of variant under investigation B.1.617 with sera of BBV152 vaccinees. Clin. Infect. Dis. ciab411. https://doi.org/10.1093/cid/ciab411.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. Curr. Protoc. Bioinformatics *69*, e96.

Zhao, S., Lou, J., Cao, L., Zheng, H., Chong, M.K.C., Chen, Z., Chan, R.W.Y., Zee, B.C.Y., Chan, P.K.S., and Wang, M.H. (2021a). Quantifying the transmission advantage associated with N501Y substitution of SARS-CoV-2 in the United Kingdom: An early data-driven analysis. J. Travel Med. *28*, taab011.

Zhao, S., Lou, J., Chong, M.K.C., Cao, L., Zheng, H., Chen, Z., Chan, R.W.Y., Zee, B.C.Y., Chan, P.K.S., and Wang, M.H. (2021b). Inferring the Association between the Risk of COVID-19 Case Fatality and N501Y Substitution in SARS-CoV-2. Viruses *13*, 638.

Zhou, Z.-Y., Liu, H., Zhang, Y.-D., Wu, Y.-Q., Peng, M.-S., Li, A., Irwin, D.M., Li, H., Lu, J., Bao, Y., et al. (2020). Worldwide tracing of mutations and the evolutionary dynamics of SARS-CoV-2. bioRxiv, 2020.08.07.242263.

Zhu, L., and Bustamante, C.D. (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics *170*, 1411–1421.

Zhu, Z., Meng, K., Liu, G., and Meng, G. (2021a). A database resource and online analysis tools for coronaviruses on a historical and global scale. Database (Oxford) *2020*, baaa070.

Zhu, Z., Liu, G., Meng, K., Yang, L., Liu, D., and Meng, G. (2021b). Rapid Spread of Mutant Alleles in Worldwide SARS-CoV-2 Strains Revealed by Genome-Wide Single Nucleotide Polymorphism and Variation Analysis. Genome Biol. Evol. *13*, evab015.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E. coli strain Top10	ThermoFisher Scientific	Cat#C404006
Chemicals, Peptides, and Recombinant Proteins		
mMESSAGE mMACHINE T7 Transcription Kit	Thermo Fisher Scientific	Cat#AM1344
SuperScript IV First-Strand Synthesis System	Thermo Fisher Scientific	Cat#18091300
Platinum SuperFi II DNA Polymerase	Thermo Fisher Scientific	Cat#12361010
SuperScript III One-Step RT–PCR kit	Thermo Fisher Scientific	Cat#12574018
GeneJET PCR Purification kit	Thermo Fisher Scientific	Cat#K0701
RNeasy Mini Kit	QIAGEN	Cat#74104
NEB Golden Gate Assembly Kit (BsaI-HFv2)	New England Biolabs	Cat#E1601L
Esp3I restriction enzyme	New England Biolabs	Cat#R0734L
Luciferase Assay System	Promega	Cat# E1501
Experimental Models: Cell Lines		
Vero E6 cells	ATCC	CRL-1586
Calu-3 cells	ATCC	HTB-55
293T cells	ATCC	CRL-3216
Oligonucleotides		
SARS-CoV-2-F1 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
SARS-CoV-2-F2 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
SARS-CoV-2-F3 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
SARS-CoV-2-F4 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
SARS-CoV-2-F5 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
SARS-CoV-2-F6 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
SARS-CoV-2-F7 sequences	Beijing Genomics Institute	"f1-f7.pdf" in Data S1
Primers for viral sgRNA assay	Beijing Genomics Institute	Table S5A; Corman et al., 2020
Primers for viral genomic RNA assay	Beijing Genomics Institute	Table S5B; Dagotto et al., 2021
Primer for competition assay	Beijing Genomics Institute	Table S5C
Recombinant DNA		
pCC1	Dr. Yonghui Zheng	N/A
pUC57	Dr. Yonghui Zheng	N/A
Software and Algorithms		
GraphPad Prism 8	GraphPad	https://www.graphpad.com
QSV analyzer	Insilicase	http://www.insilicase.com
Perl	The Perl Foundation	https://www.perl.org
BioPerl PopGen library	(Stajich, 2007)	https://bioperl.org
MUSCLE	(Edgar, 2004)	http://www.drive5.com/muscle
Bowtie2	(Langdon, 2015)	http://bowtie-bio.sourceforge.net/bowtie2
GATK Mark Duplicates	(DePristo et al., 2011; McKenna et al., 2010)	https://gatk.broadinstitute.org
SAMtools	(Li, 2011)	http://www.htslib.org
VariScan 2.0	(Hutter et al., 2006; Vilella et al., 2005)	http://www.ub.edu/softevol/variscan
SweepFineder2	(DeGiorgio et al., 2016)	http://degiorgiogroup.fau.edu/sf2.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R	The R Foundation for Statistical Computing	http://www.R-project.org
R packages: gdata (version 2.18.0), ggplot2 (version 3.3.3), RColorBrewer (version 1.1.2), scales (version 1.1.1), ggsci (version 2.9), RColorBrewer (version 1.0.12), PerformanceAnalytics (version 2.0.4), corrplot (version 0.84), lme4 (version 1.1.27.1), cgam (version 1.16), gam (version 1.20), dplyr (version 1.0.6), trend (version 1.1.4), ggtree (version 3.0.2), ape (version 5.5), ape (version 5.5), treeio (version 1.16.1), patchwork (version 1.1.1), ggtreeExtra (version 1.1.3), ggnewscale (version 0.4.4)	The R Foundation for Statistical Computing	https://cran.r-project.org
FastTree	(Price et al., 2010)	http://microbesonline.org/fasttree
GIF Movie Gear	Service Mark of CompuServe Inc	https://www.gamani.com
TreeCluster	(Balaban et al., 2019)	https://github.com/niemasd/TreeCluster
BEAST v1.10.4	(Suchard et al., 2018)	https://beast.community
Tracer	(Rambaut et al., 2018)	https://beast.community
BEAST2 PhyDyn	(Bouckaert et al., 2019; Volz and Siveroni, 2018)	https://github.com/mrc-ide/PhyDyn
UCSF Chimera	Resource for Biocomputing, Visualization, and Informatics, University of California	http://www.cgl.ucsf.edu/chimera
APBS	(Jurrus et al., 2018)	https://github.com/Electrostatics/apbs
PopArt	(Leigh and Bryant, 2015)	http://popart.otago.ac.nz/
Other		
GISAID	Freunde von GISAID e.V.	https://www.gisaid.org
CoVdb	(Zhu et al., 2021a)	http://covdb.popgenetics.net
The epidemic data of COVID-19	Global Change Data Lab	https://ourworldindata.org/covid-deaths
Nextstrain	(Hadfield et al., 2018)	https://nextstrain.org
ExPASy Server	Swiss Bioinformatics Resource Portal	https://www.expasy.org
PRJEB37886	NCBI	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB37886

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Zhenglin Zhu (zhuzl@cqu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All sequence data used here are available from GISAID (www.gisaid.org) and CoVdb (covdb.popgenetics.net). The user agreement for GISAID does not permit redistribution of sequences.

Additional Supplemental Items are available from Mendeley Data at https://doi.org/10.17632/p4mptgb9ch.2. The items also include the scripts to perform *in silico* analyses referred in this work.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animal husbandry and infection

Female golden Syrian hamsters, aged 3-4 weeks old, were obtained from Vital River Laboratories (Beijing, China). All animal experimental procedures were approved by the Animal Ethics Commission of the School of Life Sciences, Chongqing University. Hamsters were anaesthetized with isoflurane and infected intranasally with 2×10^4 PFU R203/G204 or 203K/204R virus. The hamster studies were performed twice independently. In each independent experiment, 12 hamsters received wild-type R203/G204 virus, 12 received mutant 203K/204R virus, 12 received a 1:1 mixture of R203/G204 and 203K/204R virus, and 12 received PBS (Mock). Infected hamsters were weighed and documented daily. On days 4 and 7 after infection, cohorts of 6 infected hamsters were anaesthetized with isoflurane, and nasal washes were collected using sterile DPBS. Hamsters were humanely euthanized immediately after nasal wash, and then the trachea and the four lobes of the right lung were collected as previously described (Plante et al., 2021). All hamster operations were performed under anesthesia by isoflurane to minimize animal suffering.

Cell culture and infection

African green monkey kidney epithelial Vero E6 cells (ATCC) and human lung adenocarcinoma epithelial Calu-3 cells (ATCC) were maintained in high-glucose Dulbecco's modified Eagle's medium (DMEM, GIBCO) supplemented with 10% FBS (GIBCO) at 37 °C with 5% CO_2. Cells were infected at a multiplicity of infection (MOI) of 0.01 for the indicated times. All experiments with live virus were performed under biosafety level 3 (BSL3+) conditions.

METHOD DETAILS

Identification and statistics of SARS-CoV-2 mutations

A total of 884736 full-length SARS-CoV-2 genomes were downloaded from GISAID (www.gisaid.org), NCBI and CoVdb (Zhu et al., 2021a). The collection dates of these samples ranged from December 2019 to March 2021. We assumed the sequence of strain MN908947, collected in December 2019 (Ralph et al., 2020), to represent the ancestral state of the SARS-CoV-2 genome and performed genomic alignments between the 884736 SARS-CoV-2 strains and MN908947 by using MUSCLE (Edgar, 2004). Based on these alignments, we identified mutations and performed statistics of the monthly IF with a series of Perl scripts used previously (Zhu et al., 2021b). We included the samples collected in both December 2019 and in January 2020 in a single month, January 2020 for convenience in calculating statistics. We filtered out mutations with an IF higher than 0.05 in one month and finally identified 96 mutations. Next, we performed linkage disequilibrium analysis of these mutations, according to published algorithms (Lewontin, 1964; Morton, 1955; Slatkin, 2008). We calculated ρ^2 and logarithm of odds (LOD) values (squared correlation coefficients) (Lewontin, 1964), and a statistical was applied to infer the linkage of two loci (Morton, 1955). We included mutations with $\rho^2 > 0.7$ in one linkage group. The cutoff was based on the distribution of all ρ^2 values (Figure S1A). We finally identified 12 mutation linkage groups and 26 singletons. In accordance with previous findings, we also identified LG_1 (D614G), LG_2 (L84S) and LG_3 (R203K/G204R).

We performed iSNV analyses of the raw sequence data following previous efforts (Rueca et al., 2020; Wang et al., 2020; Zhou et al., 2020; Zhu et al., 2021b). These included 14108 data from downloadable samples from the United Kingdom (PRJEB37886). We processed reads by using Bowtie2 (Langdon, 2015), GATK Mark Duplicates (DePristo et al., 2011; McKenna et al., 2010) and SAMtools (Li, 2011), requiring a mapping quality > 20 (Li, 2011).

Analyses of the confounding effects of 203K/204R and other mutants

To determine whether the increase in the IF of 203K/204R resulted from genetic hitchhiking driven by 614G (LG_1), we analyzed the IF tracks of four mutation combinations, D614/R203, D614/203K, 614G/R203 and 614G/203K. We found that 203K showed an IF increase not only in D614 but also in 614G virus (Figures S2A–S2D). The IF track of 614G/203K (Figure S2B) presents a high correlation (1, P value < 2.2e-16) with the IF track of 203K/204R (Figure 2B). There is no linkage disequilibrium between D614G and R203K/G204R (ρ^2 = 0.043, Table S1B), suggesting the independence in evolution between R203K/G204R and D614G.

To identify the causes of the decrease in D and the increase in I2, we performed correlation tests of IF tracks between pairs of mutants in the three time intervals (I1, D and I2) (Figure 2B, Table S2A). R203K/G204R show a significant positive correlation (r+) with C313T in I1, a significant negative correlation (r-) with LG_4 and an r+ with LG_6 in D. In I2, R203K/G204R present an r- with LG_4, an r- with S194L, an r+ with LG_5 and an r+ with S477N ("Cor_mut.pdf" in Data S1). We counted the frequencies of these six LGs/mutations in R203/G204 variants and 203K/204R variants (Figures S2E and S2F). We also counted the frequencies of R203/G204 and 203K/204R in the six LGs/mutants (Figures S2G–S2L). There is a significant correlation between the IF track of 222V (LG_4) in the R203/G204 virus and the IF track of R203/G204 at time intervals D and I2 (correlation = 0.96, P value = 2.933e-05, Figure S2E). R203/G204 are predominant in 222V (LG_4) (Figure S2G). The increase in 222V (LG_4) in R203/G204 variants may lead to a decrease in 203K/204R.

There is a sharp increase of 501Y (LG_5) in 203K/204R variants in I2. There is also a significant correlation between the change in the IF of 501Y (LG_5) in 203K/204R viruses and the change in the IF of 203K/204R in I2 (correlation = 0.99, P value = 0.00016, Figure S2F). 501Y (LG_5) was always found in a low percentage of the R203/G204 viruses (Figure S2E). Meanwhile, 203K/204R was predominant in the 501Y (LG_5) mutants (Figure S2H). The co-occurrence of 501Y and other mutations in LG_5 with 203K/204R should lead to the second increase in the IF of 203K/204R. 120F, 313T and 477N possibly emerged in 203K/204R strains and

increased in frequency in I1 but decreased thereafter along with the decrease in the IF of 203K/204R (Figures S2E, S2F, and S2I–S2K). The combination of these mutations and 203K/204R may result in lower fitness than the combination of 222V and R203/G204. 120F, 313T, 477N and 194L show lower IFs (< 30%) over time. Their effects on 203K/204R are negligible.

There are $2^3 = 8$ possible combinations of three sets of two-allele polymorphisms. To further understand the confounding effects of A222V (LG_4) and N501Y (LG_5) on 203K/204R, we evaluated the IF changes of the 8 combinations of these polymorphisms and identified four dominant lineages, ANR, VNR, ANK and AYK (Figure 3A). The global IF changes are consistent with the median IFs in different countries (Correlations > 0.89, P value < 0.05, Figures S3A–S3H). The increases in VNR occurred mostly in European countries (Figure S3K). In the UK (Figure 3B), the increase in VNR was accompanied by decreases in ANR and ANK. With the subsequent increase in AYK, VNR, ANR and ANK vanished (Figure 3B). In Mexico and India (Figures 3C and 3D), there were continuous increases in ANK and decreases in ANR. ANK and AYK show a slower spread in North American countries (e.g., the USA) (Figures 3E, S3I, and S3J) than in European countries (e.g., the UK) (Wilcoxon test, P value = 0.00017). In South Africa, AYR (A222 + 501Y + R203/G204) spread rapidly and finally replaced the preceding ANR and ANK lineages (Figure 3F). The AYR variants mostly belong to the B.1.351 (Beta) (Mwenda et al., 2021) lineage (1211/1379, 87.8%).

Following a reported approach (Korber et al., 2020), we compared the growth rates of these lineages separately in multiple geographically restricted contexts, with selected time windows in the temporal vicinity of the introduction of a new variant. In four hierarchical geographic levels (1, worldwide; 2, continent; 3, country; 4, state/city), we counted the weekly running counts (the folder "Weekly Running Counts" in Data S1) and performed Fisher's exact test of the fraction of pairs of lineages on the onset (the first day when there are more than 15 cumulative samples for both lineages) and the day after more than two weeks for the three time intervals (ANR versus ANK in I1, ANK versus VNR in D and VNR versus AYK in I2). Maximum likelihood estimation of fraction trends was constructed (Figure 3G and the full result data is in the folder "Maximum Likelihood Estimation" of Data S1) and the trends were evaluated by Mann-Kendall trend test and by isotonic regression analysis. We tested the null hypothesis that the growth rates of both lineages are equal. From a perspective of all hierarchical geographic level, it was found that ANK had a higher growth rate than ANR in I1, and so did VNR when compared to ANK in D, as well as AYK when compared to VNR in I2 (Tables S2C and S2D). Although there are a few exceptional cases (Table S2D) that possibly due to the impact of other mutants, genetic drift or limited sample size, our results confirmed the adaption order of these four lineages.

In details, we compared the growth rates of three pairs of lineages in three time intervals, respectively. We calculated and plotted the weekly running counts of pairs of lineages by the R package "ggplot2," based on which the fraction of lineages in the onset day (the first time point) and the day more than 14 days later (the second time point) were counted. It is required that the counts are higher than 15 for both lineages in the two time points and such case repeated for more than 3 times in the proximal days. We performed simulation and plotted the maxium likelihood estimation of lineage fractions by the R packages "lme4," "cgam" and "ggplot2," requiring a count > 5 for both lineages in more than 14 sampling days (Korber et al., 2020). We performed randomization of the data for 1000 times and ranking test by comparing the null hypothesis with fraction increase/decrease. The trend of lineage fractions was evaluated by Mann-Kendall trend test, requiring a statistical significance with a < 0.05 P value for both tests referred above. Binomial test was performed to evaluate the overall significance for fraction increased or decreased cases in difference geographic levels (Tables S2C and S2D) at the end.

Evolutionary analysis

We built a multiple sequence alignment file by retrieving sequences from pairwise sequence alignments using MN908947 as the reference. We performed sliding window analysis with a window size of 200 bp and a step size of 50 bp. We calculated Pi, Theta (Tajima, 1993) and Tajima's D (Tajima, 1989) values by using VariScan 2.0 (Hutter et al., 2006; Vilella et al., 2005), and the composite likelihood ratio (CLR, step size = 50) (Nielsen et al., 2005; Zhu and Bustamante, 2005) by using SweepFineder2 (DeGiorgio et al., 2016). To test whether a single mutation site was under positive selection, we compared the median of the CLR in the 100 bp nearby region of the mutation (m) and the top 5% threshold (t) of all CLR values. We used $CLR_{m/t} = m/t$ to infer the significance of peaks. A $CLR_{m/t} > 1$ implies a significant CLR peak and provides evidence of positive selection. We calculated $CLR_{m/t}$ for a 3000 bp region centered on 28881-28883 (the location of R203K/G204R). We performed the calculations for R203/G204 and 203K/204R variants, and then conducted a comparison.

We extracted the nucleotides at the 96 mutation sites from all SARS-CoV-2 genomes and assembled the nucleotides into consecutive sequences; i.e., we used a shortened sequence (SS) with full mutation information to represent the whole genome. We clustered these SSs by month and by allele. For comparisons of Pi, Theta and Tajima's D between R203/G204 and 203K/204R variants, we used the BioPerl PopGen library (Stajich, 2007) to perform population genetic calculations within R203/G204 SSs and 203K/204R SSs, respectively.

We used SS to construct the phylogenetic tree of SARS-CoV-2. For a better clarification of the relationships between strains in the phylogenetic tree, we discarded duplicate SSs obtained within the same collection month and country. We finally obtained a unique SS dataset with 322893 sequences. We built a multiple alignment file based on the alignments of these SSs and the reference. Then, we constructed a phylogenetic tree by using FastTree (Price et al., 2010). We used the R script library ggtree (Yu, 2020) to annotate and color the phylogenetic tree according to lineages, collection months or collection regions. We used the software GIF Movie Gear (www.gamani.com) to build animations to show the changes in IF in different countries over time.

We followed a reported pipeline (Volz et al., 2021) to evaluate the transmission advantages of 203K/204R compared to R203/G204. Generally, we constructed a phylogenetic tree using all SARS-CoV-2 sequences collected in the UK from December, 2019 to July

15th. The phylogenetic clusters were identified by TreeCluster (Balaban et al., 2019) with the parameter -t 0.045. We performed simulations in a logistic growth model by using BEAST v1.10.4 (Suchard et al., 2018) with an HYK substitution model, a strict clock type and an exponential prior distribution. We set the lengths of chains to 10e7 for cocirculating sequences and 5e5 for all sequences. We used PhyDyn in BEAST2 to perform simulation with the SEIR model (Bouckaert et al., 2019; Volz and Siveroni, 2018) and followed the phylogeographic model provided by (Volz et al., 2021). The plotting and analysis of the simulation data were performed with R.

Function prediction based on clinical and epidemic data

We manually gathered the clinical information of 36750 SARS-CoV-2 strains from GISAID. As in our previous work (Zhu et al., 2021b), we grouped this information into four pairs of opposite patient statuses according to a series of keywords (Table S4A). We counted the ratios of variants with different patient statuses and tested the significance by using the chi-square test.

We calculated CFRs in different countries using the epidemic data provided by ourworldindata.org/covid-deaths. We performed a correlation test between the median CFRs and the IFs of mutations among different months. Because the mean duration from the onset of symptoms to death in COVID-19 patients is 17 or 18 days (Verity et al., 2020), we performed an x+1 correlation test, in which an IF in month x was compared with a CFR in month x+1. We evaluated the significance of the correlation through a two sided test.

Generation of the R203K/G204R mutant virus

The SARS-CoV-2 virus was generated by using a reverse genetic method (Plante et al., 2021; Xie et al., 2021; Xie et al., 2020). Seven different DNA fragments spanning the entire genome of SARS-CoV-2 (USA_WA1/2020 SARS-CoV-2 sequence, GenBank accession number MT020880) were synthesized by Beijing Genomics Institute (BGI, Shanghai, China) and cloned into the pUC57 or pCC1 (kindly provided by Dr. Yonghui Zheng) plasmid by standard molecular cloning methods. The sequences of the F1∼F7 fragments and the restriction enzymes used for digestion and ligation are listed in the file "f1-f7.pdf" in Data S1. Full-length cDNA assembly and recombinant SARS-CoV-2 virus recovery were performed as previously described (Plante et al., 2021; Xie et al., 2021; Xie et al., 2020). Briefly, the full-length cDNA of SARS-CoV-2 was assembled via the *in vitro* ligation of contiguous cDNA fragments. Then, full-length genomic RNA was collected by *in vitro* transcription and electroporated into Vero E6 cells. The SARS-CoV-2 virus was harvested at 40 h post electroporation, and viral titers were determined by plaque assays. For the generation of the R203K/G204R mutant virus, GGG®AAC nucleotide substitutions were introduced into a subclone of pUC57-F7 containing the nucleocapsid gene of the SARS-CoV-2 wild-type infectious clone by overlap-extension PCR. The primers are shown in Figure S6S.

Plaque assays

Approximately 1 × 10^6 cells were seeded into each well of 6-well plates and cultured at 37 °C under 5% CO_2 for 12 h. R203/G204 or 203K/204R viruses were serially diluted in DMEM with 2% FBS, and 200 μL aliquots were transferred to cell monolayers. The viruses were incubated with the cells for 1 h. After incubation, overlay medium was added to the infected cells in each well. The overlay medium contained DMEM with 2% FBS and 1% sea-plaque agarose. After 2 days of incubation, the plates were stained with neutral red, and plaques were counted in a light box.

Neutralization assay

Neutralization assays were performed using R203/G204 and 203K/204R mNeonGreen viruses as previously described (Plante et al., 2021). In brief, Vero cells were plated in 96-well plates. On the following day, sera were serially diluted and incubated with an R203/G204 or 203K/204R mNeonGreen virus at 37 °C for 1 h. The virus-serum mixtures were transferred to a Vero cell plate at a final MOI of 2.0. After 20 h, Hoechst 33342 solution was added to stain the cell nuclei, the cells were sealed with a membrane and incubated at 37 °C for 20 min, and mNeonGreen fluorescence was quantified. The total numbers of cells (indicated by nucleus staining) and mNeonGreen-positive cells were quantified in each well. Infection rates were determined by dividing the mNeonGreen-positive cell number by the total cell number. Relative infection rates were obtained by normalizing the infection rates of serum-treated groups to those of non-serum-treated controls. A nonlinear regression method was used to determine the fold dilution that neutralized 50% of mNeonGreen fluorescence (NT_{50}). The curves of the relative infection rates versus the serum dilutions (log_{10} values) were plotted using GraphPad Prism 8.

Viral infection in a primary human airway tissue model

For the assessment of viral replication kinetics, either an R203/G204 or 203K/204R virus was inoculated into a culture at an MOI of 5 in PBS. After 2 h of infection at 37 °C with 5% CO2, the inoculum was removed, and the culture was washed three times with PBS. The infected epithelial cells were maintained without any medium in the apical well, and medium was provided to the culture through the basal well. The infected cells were incubated at 37 °C under 5% CO2. From day 1 to day 5, 300 μL PBS was added on the apical side of the airway culture, and the culture was incubated at 37 °C for 30 min to elute the released viruses.

Validation of competition assay

Ratios of R203/G204: 203K/204R RNA were determined via RT-PCR with quantification of Sanger peak heights. Briefly, R203/G204 and 203K/204R viruses were mixed at PFU ratios of 1:1, 3:1 and 9:1 based on their PFU titers. To quantify R203/G204: 203K/204R ratios, a 596 bp RT-PCR product (Primers: SARS-CoV-2 28354F, 5¢-CCAGAATGGAGAACGCAGTG-3¢; SARS-CoV-2 28949R, 5¢-TGTCAAGCAGCAGCAAAGC-3¢) was amplified from the extracted RNA using a SuperScript III One-Step RT–PCR kit (Thermo Fisher

Scientific) according to the manufacturer's instructions. The PCR product was purified by a GeneJET PCR Purification kit (Thermo Fisher Scientific) and submitted to Sanger sequencing (BGI, Shanghai, China) (Primer for Sanger sequencing: 5¢-CCAGAATGGA GAACGCAGTG-3¢). The sequence electropherograms were further scored by QSV analyzer to quantify the proportion of R203/G204 and 203K/204R viruses. The correlation between input PFU ratios and output RT–PCR amplicon ratios and verification of the actual ratios of R203/G204: 203K/204R achieved upon viral mixing are shown in Figures S6A and S6B.

Viral sgRNA assay and genomic RNA assay

Viral sgRNA assay was performed by using a leader-specific forward primer, as well as a reverse primer and a probe targeting envelope protein (E) gene sequences, as previously described (Corman et al., 2020; Wölfel et al., 2020; Dagotto et al., 2021). Infectious cell lysates were harvested at indicated time points and then total RNA was extracted using an RNeasy Mini Kit (QIAGEN, Hilden, Germany). RT-PCR was performed with a SuperScript III One-Step RT–PCR kit (Thermo Fisher Scientific) and an ABI StepOnePlus PCR system (Applied Biosystems, CA, USA), according to the manufacturer's instructions. Thermal cycling was performed at 50°C for 10 min for reverse transcription, followed by 95°C for 3 min and then 45 cycles of 95°C for 15 s, 58°C for 30 s. The oligonucleotide sequences of the primers are as previously described (Corman et al., 2020). E_Sarbeco_F: 5¢-ACAGGTACGTTAATAGTTAA TAGCGT-3¢; E_Sarbeco_R: 5¢-ATATTGCAGCAGTACGCACACA-3¢; E_Sarbeco_P1 (Probe): 5¢-FAM-ACACTAGCCATCCT TACTGCGCTTCG-BHQ1-3¢.

We used Orf1ab gene as the target site of primers for the quantification of genomic RNA, since Orf1ab does not generate subgenomic transcripts (Kim et al., 2020). The oligonucleotide sequences of the primers are as previously described (Dagotto et al., 2021). SARS-CoV2.ORF1ab.F: 5¢-GGCCAATTCTGCTGTCAAATTA-3¢; SARS-CoV2.ORF1ab.R: 5¢-CAGTGCAAGCAGTTTGTGTAG-3¢; SARS-CoV2.ORF1ab.P (Probe): 5¢-FAM-ACAGATGTCTTGTGCTGCCGGTA-BHQ1-3¢.

Pathological examination

At indicated time points, hamsters were anaesthetized with isoflurane and lung lobes were harvested. Tissues fixed in 10% formalin were trimmed and embedded in paraffin. The paraffin blocks were cut into 3 μm-thick sections and were stained using a standard hematoxylin and eosin procedure. The histopathology scoring of infected hamsters was based on the percentage of inflammation area in each section of the lung lobes collected from each animal, using a semiquantitative pathology scoring system adapted from (Hou et al., 2020). Lung lobes were scored individually using the following scoring system: 0, no pathological change; 1, affected area (£10%); 2, affected area (> 10% and £30%); 3, affected area (> 30% and £50%); and 4, affected area (> 50%). We obtained 5 slices from each hamster, the scores of lung slices were added to determine the total pathology score per animal. All scoring was performed by the same operator to ensure scoring consistency.

3D structural model of assembled nucleocapsid

Atomic models (PDB accession code 6M3M, 6WZQ) of the nucleocapsid protein NTD and CTD were fitted to MHV RNP densities (Gui et al., 2017) using the "fit to segments" tool in UCSF Chimera (Gui et al., 2017). The theoretical isoelectric point (pI) of each region was predicted with the Compute pI/Mw tool on the ExPASy Server (Gui et al., 2017). The electrostatic surface potential was generated with APBS (Jurrus et al., 2018) and viewed in UCSF Chimera. The cryo-EM electron density map of MHV RNP was kindly provided by Dr. Xiang Y at Tsinghua University.

Haplotype network

We first created 7-bp sequences by combining the nucleotide sequences of different strains at seven mutation sites (R203K/G204R (LG_3), N501Y (LG_5), A222V (LG_4), C313T, S477N, S194L and I120F). We then calculated the IFs and geological distribution of these 7-bp sequences. With these data, we plotted the haplotype network by using PopArt (Leigh and Bryant, 2015).

QUANTITATIVE AND STATISTICAL ANALYSIS

We used the R package "corrplot" to perform correlation tests between pairs of LGs/mutations. Mann-Kendall trend test was performed by the R package "trend." Wilcoxon test, Binomial test and Fisher's exact test were performed by R. We wrote Perl scripts to classify the strains into lineages and quantified the IFs of these lineages. The source data are from GISAID and Nextstrain (Hadfield et al., 2018). Heatmaps, box-plots and scatter-plots were generated using the R libraries "gdata" and "ggplot2."

Cell Host & Microbe, Volume *29*

Supplemental information

Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2

Haibo Wu, Na Xing, Kaiwen Meng, Beibei Fu, Weiwei Xue, Pan Dong, Wanyan Tang, Yang Xiao, Gexin Liu, Haitao Luo, Wenzhuang Zhu, Xiaoyuan Lin, Geng Meng, and Zhenglin Zhu

Supplementary Figures

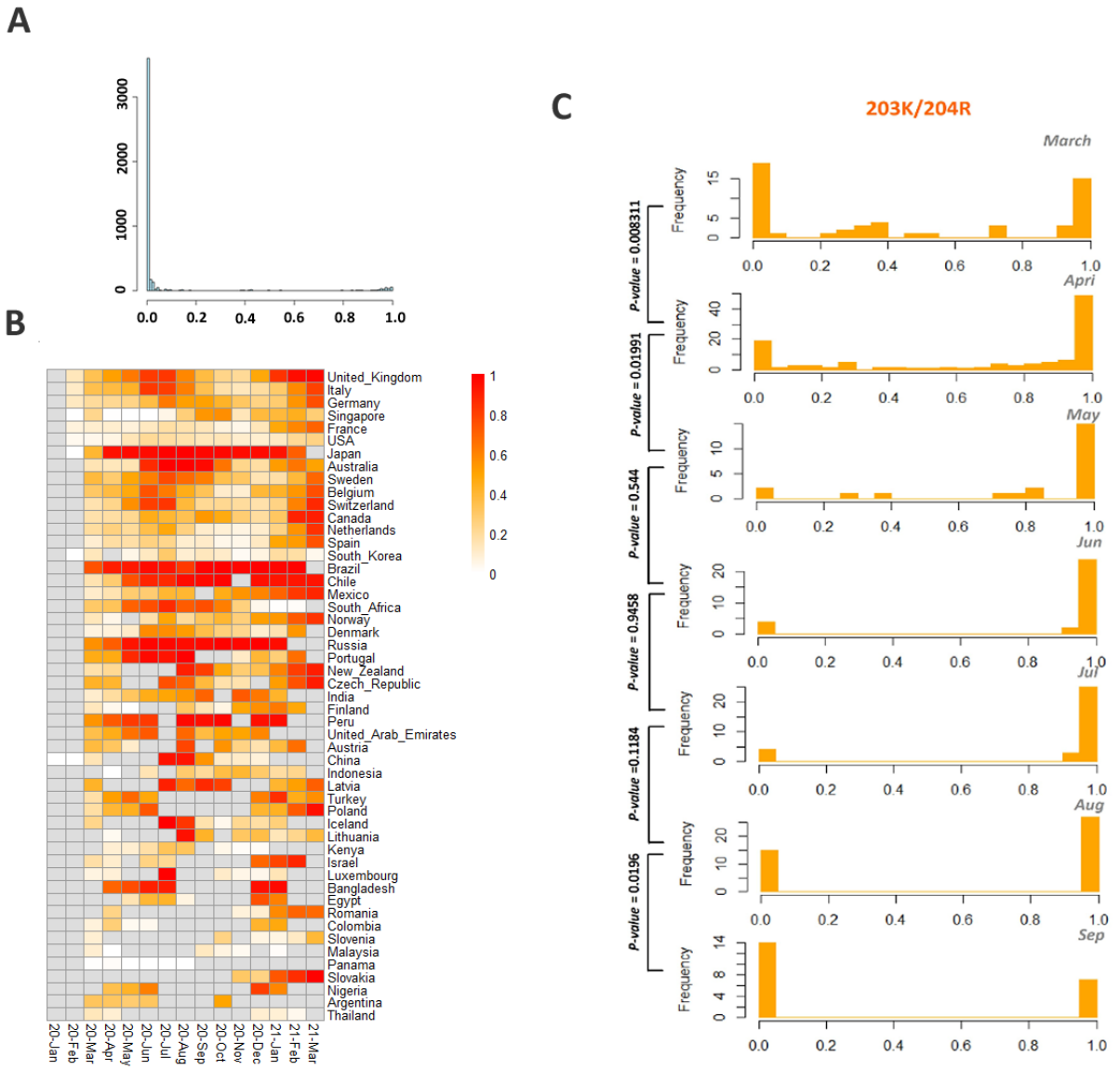


Figure S1. Statistics of SARS-CoV-2 mutations, related to Figure 2.

(A) The distribution of ρ^2 for pairs of mutations. (B) A heatmap showing the IF changes in different countries over months. The data corresponds to Figure 2C. We did not show the countries without an IF > 0 identified in any month, mostly due to an insufficient number of samples. The empty IFs are coloured grey. (C) The changes in the distribution of IFs over time for 203K/204R. The iSNV analyses are based on SARS-CoV-2 raw-sequence data collected in the United Kingdom. The X-axis represents the IF, whose value ranges from 0 to 1 (fixed), and the Y-axis refers to the counts of samples with different IF values. The results of the Wilcoxon test for the difference between two distributions are shown at the left of each figure.

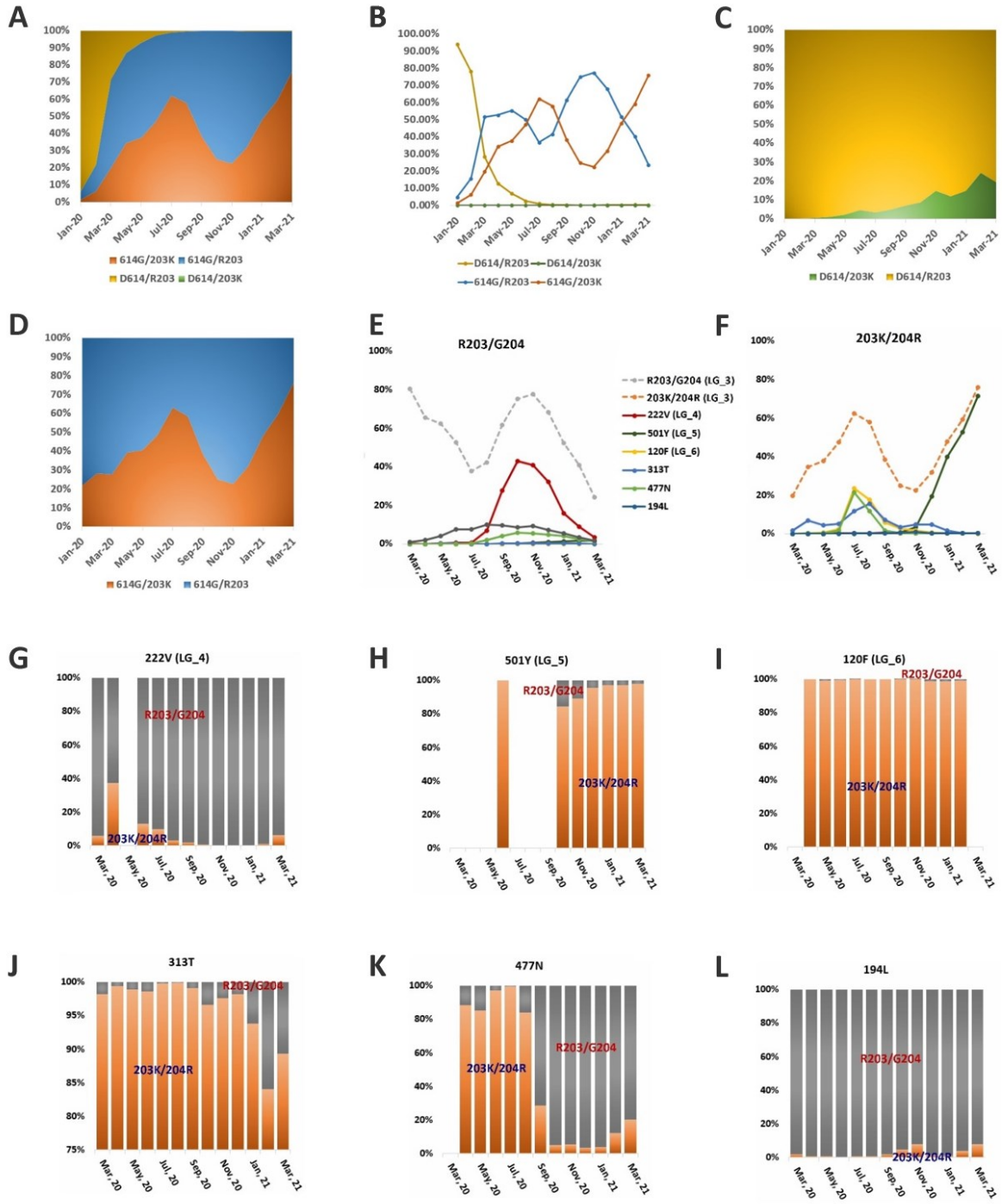


Figure S2. Competition and cooperation of 203K/204R and other mutants, related to Figure 3.

(A-D) The changes in the IFs of four mutation combinations, D614/R203, D614/203K, 614G/R203 and 614G/203K. (A, C and D) are the percentage accumulated area maps. (A) is for four mutation combinations. (C and D) are for D614/R203 vs D614/203K and 614G/R203 vs 614G/203K, respectively. (B) is the line chart of the IFs according to (A). (E-L) are the comparison of IF tracks for the six mutants with a significant correlation with LG_3. (E and F) are the IFs of the six mutants (real lines) in the R203/G204 and 203K/204R variants, respectively. In (E) and (F), dotted lines denote the changes in the IFs of the R203/G204 and 203K/204R variants, respectively. (G-L) are the ratios of R203/G204 and 203K/204R in the six mutants in different months.

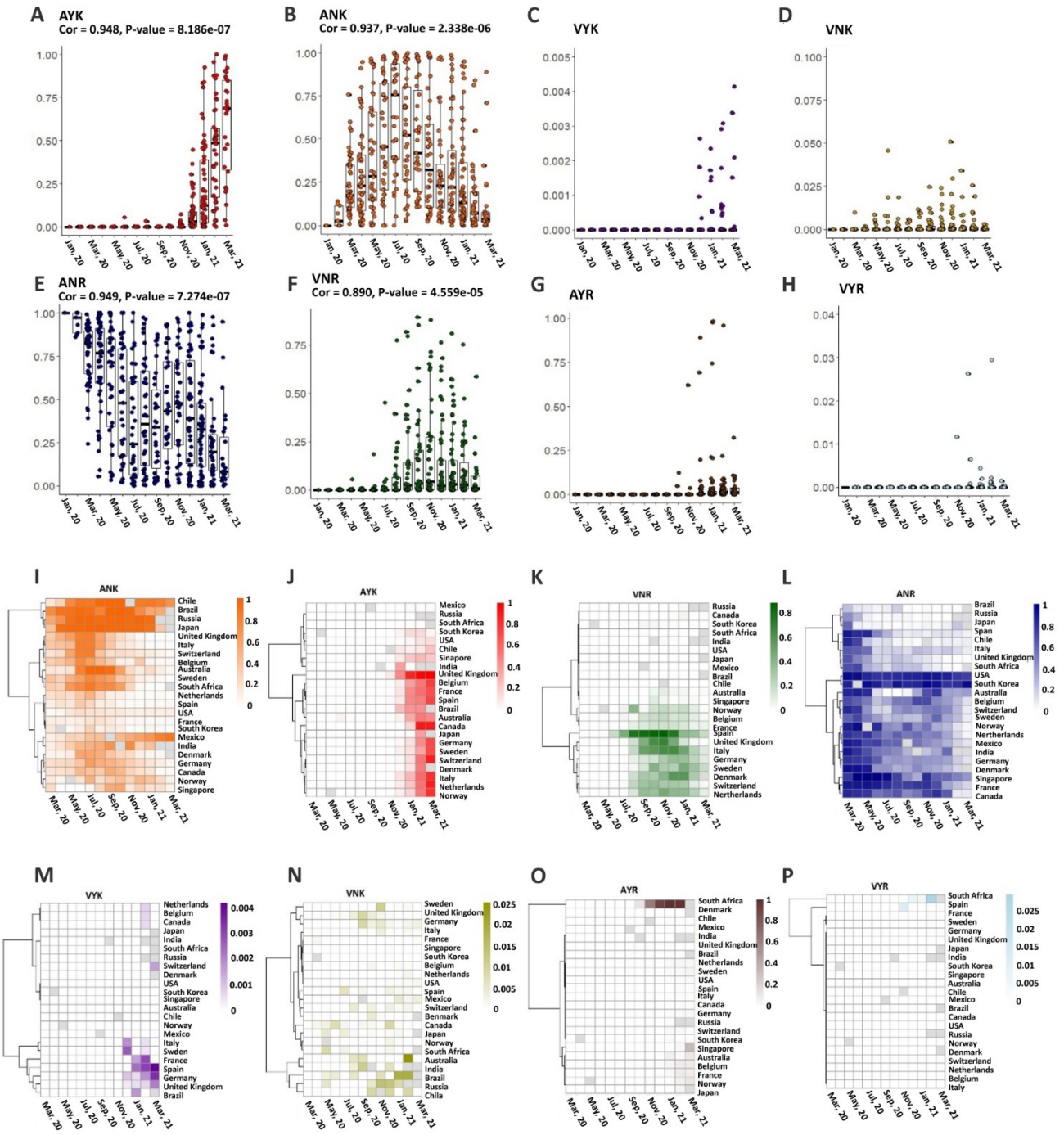


Figure S3. IFs of lineages in different countries, related to Figure 3.

(A-H) The IFs in different countries for the 8 lineages shown in Figure 3. For the four dominant lineages shown in Figure 3A, we calculated the correlations between the median IFs in different countries and the worldwide IFs shown in Figure 3A. The results are displayed at the top. (I-P) The IF changes in different countries for the eight lineages shown in Figure 3. We did not show the countries without an IF > 0 identified in any month, mostly due to an insufficient number of samples.

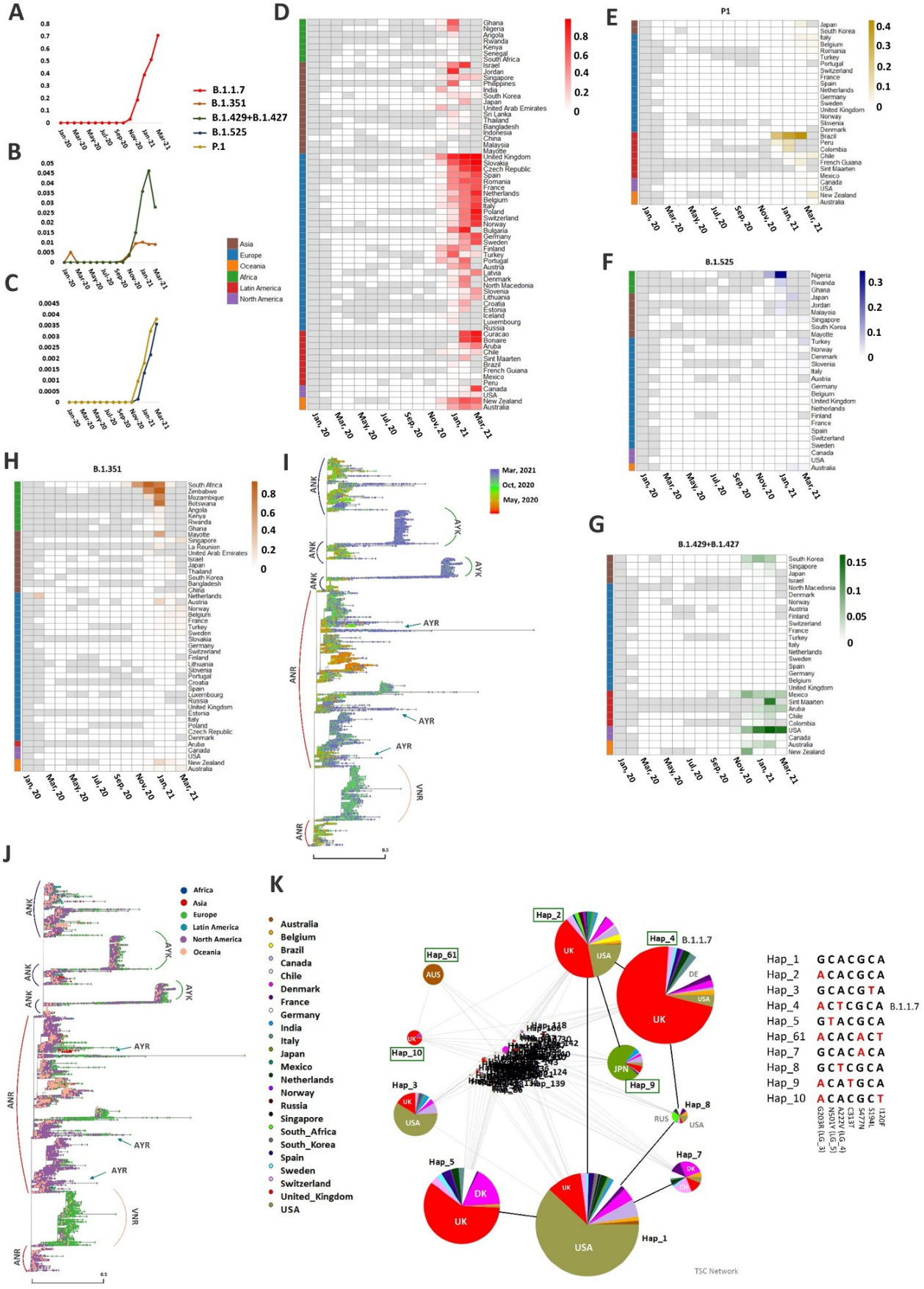


Figure S4. Analyses of R203K/G204R related lineages, related to Figure 3.

(A-H) The changes in the IFs of fast-spreading lineages. (A-C) The changes in the IFs of five lineages around the world. (D-H) The changes in the IFs of lineages in different countries. Continents are marked in the annotation row at the left. (I and J) The distributions of the collection months and continents in the SARS-CoV-2 phylogenetic tree. In (I), the colours of the tip nodes denote the collection months. In (J), the colours of tip nodes denote the collection continents. Lineages are marked. (K) A network of haplotypes constructed from 7 LGs/mutations. The geological positions are differentiated by different colours. The sequences of the haplotypes are shown on the right, with mutants marked in red.

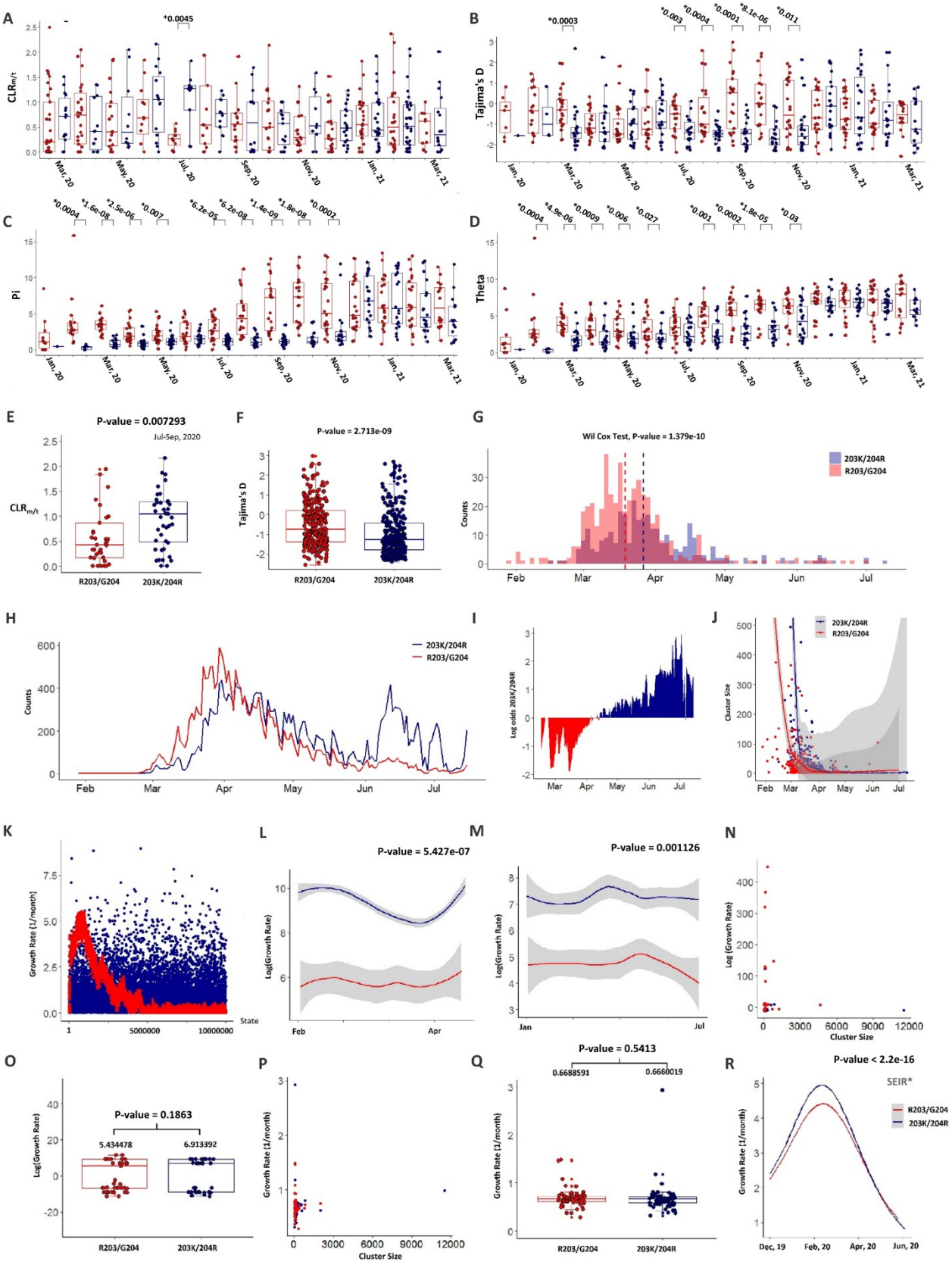


Figure S5. Selection signatures for R203K/G204R, related to Figure 4.

(A) Comparison of CLR $_{m/t}$ between R203/G204 (red) and 203K/204R variants (dark blue) in different months. (B-D)

Comparisons of Tajima's D (B), Pi (C) and Theta (D) in the whole genome of R203/G204 (coloured in red) and 203K/204R strains (coloured in dark blue). Comparisons with statistical significance are marked. (E) Comparison of CLR $_{m/t}$ in countries and in the months from July to September between R203/G204 (red) and 203K/204R variants (dark blue). (F) Comparison of Tajima's D calculated by month and country between R203/G204 (red) and 203K/204R viruses (dark blue). (G-J) The temporal distribution of R203/G204 and 203K/204R phylogenetic clusters in the United Kingdom. (G) Counts of UK R203/G204 and 203K/204R clusters when first detected and over time. (H) Numbers of collected R203/G204 or 203K/204R samples over time. (I) Log odds of the frequency of R203K/G204R over time. (J) Relationship between cluster size (Y-axis) and the date when the first sample was collected within a cluster (X-axis). (K-R) Comparison of the phylodynamic growth rates between R203/G204 and 203K/204R variants. (K) The growth rates of cocirculating clusters across states in a logistic growth model. (L and M) are the growth rates (logged in the plot) over time simulated in the skygrowth coalescent model for cocirculating clusters and all clusters, respectively. (N) The median growth rates of clusters (logged in the plot, Y-axis) versus cluster sizes (X-axis). (O) Comparison of the median growth rates of clusters between R203/G204 and 203K/204R. The legends in (P) follow (N) but without log transformation. Those in (Q) follow (O). (N and O) are the simulation results of the skygrowth coalescent model. (P and Q) are the simulation results of the logistic growth model. (R) is the fitted curves of the growth rates from the phylodynamic SEIR model. Both genetic and phylogenetic data are used (SEIR*).

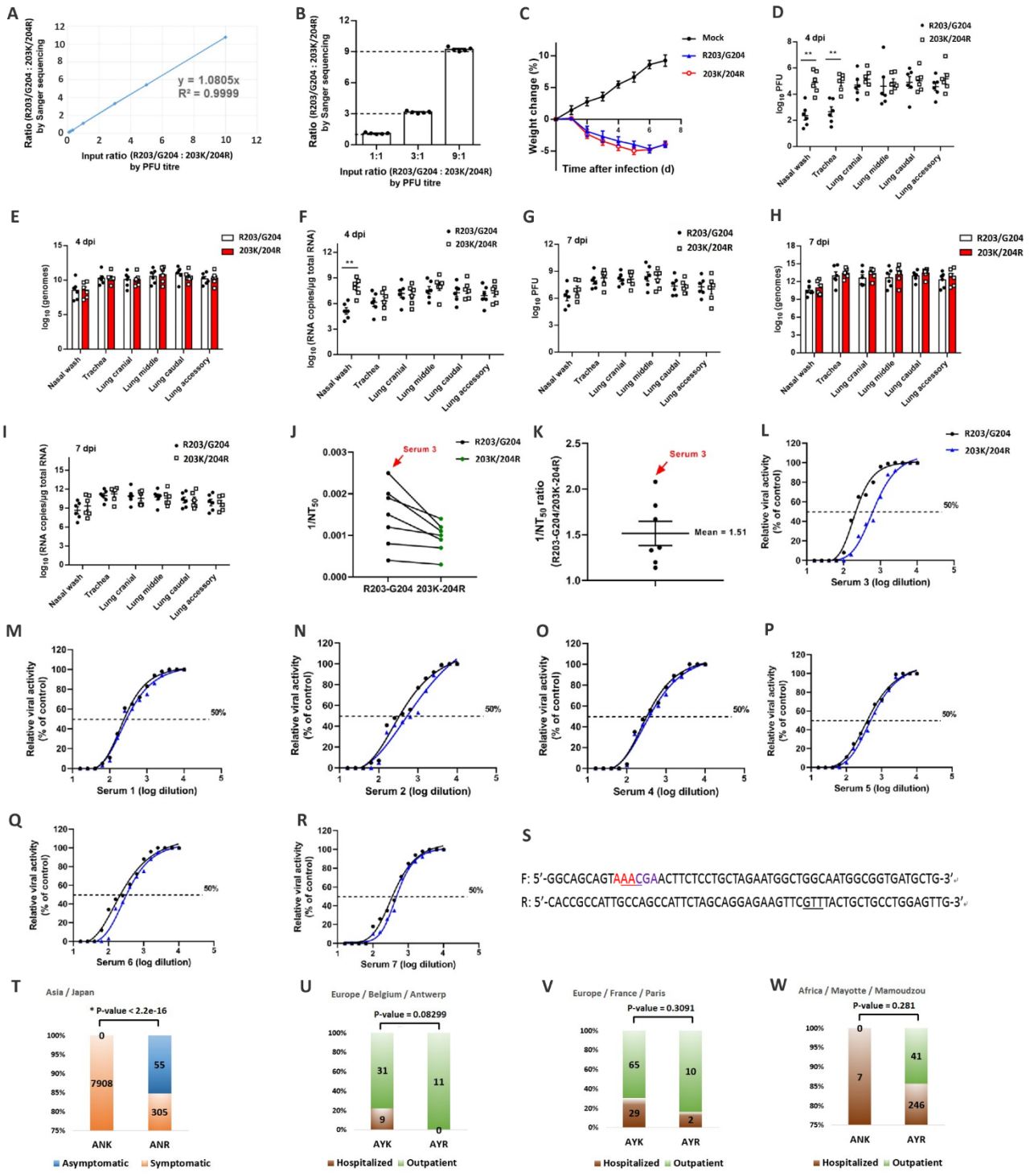


Figure S6. Experiments and analysis results focusing on the fitness, infectivity and virulence of R203K/G204R, related to Figures 4-6.

(A and B) show the validation of competition assay. (A) The correlation between input PFU ratios and output RT–PCR amplicon ratios. The relative amounts of R203/G204 and 203K/204R RNAs was quantified by RT–PCR and Sanger sequencing. R203/G204 and 203K/204R viruses were mixed at PFU ratios of 1:1, 1:3, 1:5, 1:10, 3:1, 5:1 and 10:1. To quantify R203G204/203K204R ratios, RT-PCR products were amplified from extracted RNA and submitted to Sanger sequencing. Data were analyzed by linear regression with coefficient of determination (R2). (B)

Verification of the actual ratios of R203/G204: 203K/204R achieved upon viral mixing. The mixing procedures were repeated independently for five times. The samples were collected immediately after mixing. The results showed that the R203/G204: 203K/204R ratios of intended 1:1, 3:1 and 9:1 were actually 1.06±0.04: 1, 3.21±0.13: 1 and 9.48±0.36: 1, respectively. (C-F) show an independent repeat of hamster experiments infected with R203/G204 and 203K/204R viruses. (C-F) In each independent experiment, 12 hamsters received wild-type R203/G204 virus, 12 received mutant 203K/204R virus, 12 received a 1:1 mixture of R203/G204 and 203K/204R virus, and 12 received PBS (Mock). Weight loss (C) was monitored for 7 dpi. Data are presented as mean ± s.e.m.; n = 12 (all cohorts) at days 0–4; n = 6 (all cohorts) at days 5–7. Weight loss was analysed by two-factor analysis of variance (ANOVA) with Tukey's post hoc test. Infectious titres (D) and amounts of viral genomes (E) were quantified in nasal wash, trachea and lung samples on the 4th dpi. Dots represent individual hamsters (n = 6). The E sgRNA loads (F) at 4 dpi were calculated as a measurement of infectivity. Dots represent individual hamsters (n = 6). Data are presented as the mean ± s.e.m.. **, $p<0.01$. (G-I) are the results of hamster experiments infected with R203/G204 and 203K/204R viruses at 7 dpi. Infectious titres (G) and amounts of viral genomes (H) were quantified in nasal wash, trachea and lung samples on the 7 dpi. Dots represent individual hamsters (n = 6). The E sgRNA loads (I) at 7 dpi were calculated as a measurement of infectivity. Dots represent individual hamsters (n = 6). Data are presented as the mean ± s.e.m.. (J-R) show the neutralizing activities of hamster sera against R203/G204 and 203K/204R mNeonGreen SARS-CoV-2 viruses. (J) Neutralizing activities of hamster sera against R203/G204 and 203K/204R viruses with a mNeonGreen reporter. $1/NT_{50}$ values were plotted. Symbols represent sera from individual hamsters. (K) Ratio of $1/NT_{50}$ between R203/G204 and 203K/204R viruses. Symbols represent sera from individual hamsters. (L-R) Neutralization curves of serum from individual hamsters. The solid line represents the fitted curve, and the dotted line indicates 50% viral inhibition. Data in (K) are represented as the mean ± s.e.m.. (S) The primers used for overlap-extension PCR in the generation of the R203K/G204R mutant virus. Underlined nucleotides are nucleotide substitutions. Colored nucleotides denote the reading frame (R203K/G204R mutation). (T-W) are the prediction of the clinical outcomes of ANK, ANR, AYK and AYR infections based on clinical data collected on a small scale. Legends follow Figure 6.

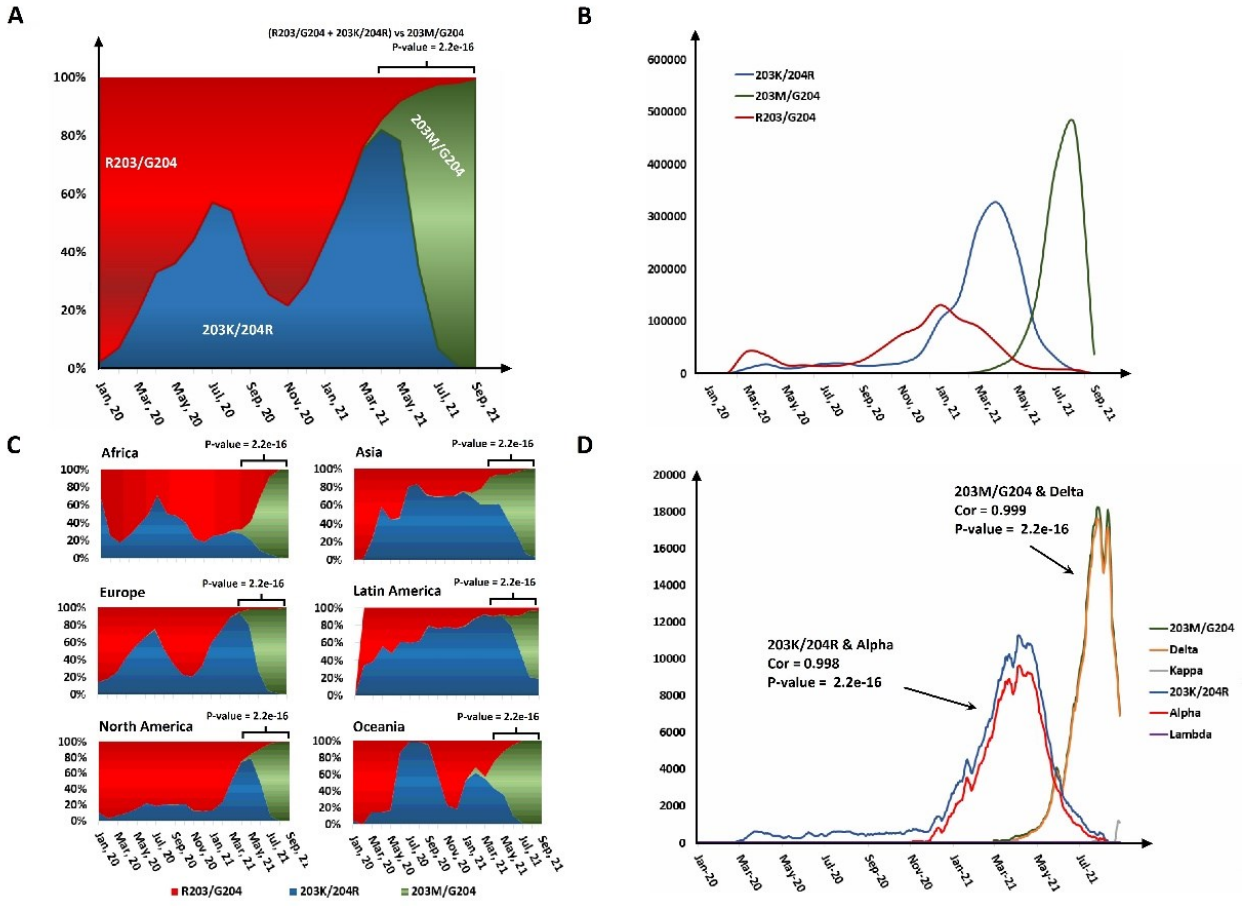


Figure S7. The IF change of mutations in 28881 to 28883, related to Figure 2.

(A) The IF change of R203/G204, 203K/204R and 203M/G204 in the world from January, 2020 to September 2021. (B) The change of the counts of the three alleles in the world up to date. (C) The IF change of the three alleles in different continents. (D) the change of the counts of 203K/204R and 203M/G204 and four main lineages with one of the two mutations. In (A and C), Fisher's exact test was performed on the fraction of R203/G204 + 203K/204R and 203M/G204 in the beginning and the recent months of emergence of R203M.