Epidemic trends of SARS-CoV-2 associated with immunity, race, and viral mutations

Yasuhiko Kamikubo[1,*], Toshio Hattori[2,3], and Atsushi Takahashi[2,3 *]

[1]Department of Human Health Sciences, Graduate School of Medicine, Kyoto University, Sakyo-ku, Kyoto, 606-8507, Japan. [2]Graduate School of Health Science Studies and [3]Research Institute of Health and Welfare, Kibi International University, Takahashi, Okayama 716-8508, Japan.

*e-mail: kamikubo.yasuhiko.7u@kyoto-u.ac.jp; atakah7@kiui.ac.jp

Abstract

The world has been plagued by complex waves of SARS-CoV-2 epidemics that vary from region to region, leaving the end of the pandemic unpredictable. Here we performed "genetic fingerprinting" to compare the local viral genotypes with epidemiological information and reveal the molecular dynamics of the SARS-CoV-2 epidemic worldwide. A multifaceted analyses of the epidemic trends and their relationship to virus genotypes, regional herd immunity, population density, and race has shown that epidemic outcomes are affected by: (1) Increased fitness of the virus due to mutations of viral proteins; (2) Immunity against previously prevalent subtypes that prevents or exacerbates COVID-19; (3) Immune evasion due to viral mutations; (4) Viral competition with coexisting subtypes; (5) Dense and crowded living environment; (6) Racial and social disparities; (7) Upper limit of viral mutations that enable natural selection. These findings provide an overview of the current epidemic, help predict the future, and develop effective countermeasures.

Key words: SARS-CoV-2, COVID-19, viral mutation, herd immunity, social disparity

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1] outbreak in Wuhan, China in early December 2019 became a pandemic with more than 1.4 million deaths reported worldwide. Containment strategies focus primarily on mobility restrictions, quarantine, and contact tracing. However, even in Western countries that have enforced strict social blockades, the death toll is high, and many countries have been hit by new epidemics when the blockade was lifted. It is difficult to grasp an overview of global epidemic trends and the future is unclear.

Epidemiology is expected to predict the future of infection, but there is too much uncertainty to narrow down from the many possibilities.[2,3] Every time humans face a crisis of survival, they have developed new tools to overcome the predicament and prosper. Previously, we developed epidemiological tools using the influenza epidemic curves.[4] Analyses of epidemic trends in Japan revealed that multiple subtypes of SARS-CoV-2 had invaded.[5] The tools also predicted epidemic trends in Europe and the United States (USA) and suggested appropriate countermeasures.[4]

Practical epidemiology of infectious diseases requires that conclusions be drawn as soon as possible to get a complete picture of the outbreak and lead to activities to prevent and end its spread. Therefore, we introduced epidemiological parameters and Fermi estimates to solve formulas that predict case fatality rate (CFR) and derived the hypothesis that herd immunity and antibody-dependent enhancement (ADE) determine the severity of coronavirus disease 2019 (COVID-19).[5] We have presented Spike: D614G as a mutation that can cause ADE, suggesting the involvement of this epitope in immune pathogenesis.[5] To our knowledge, it was the first paper focusing on Spike: D614G. Subsequent papers have revealed that this mutation also increases infectivity of the

virus.[6,7] However, while Fermi estimates are useful for predicting the future in situations where timely action is required, it is known that differences in assumptions and inference methods can cause considerable errors in conclusions. In this study, “genetic fingerprinting”[8] was adopted for the epidemiological analyses of the current state of the pandemic. The results validate the hypotheses we have proposed in the past[5] and reveal the global trends of SARS-CoV-2 epidemics.

Results

To perform "genetic fingerprinting",[8] we referred to Global Initiative on Sharing All Influenza Data (GISAID) website[9] to investigate global trends in viral genetic mutations The second wave of epidemic in Europe that began in the autumn of 2020 has been mainly due to ORF10: V30L; N: A220V/ORF14: L67F; Spike: A222V variant, probably originated in Spain[10] and spreading to Norway, Italy, Switzerland, the United Kingdom (UK), and New Zealand (Fig. 1a). GISAID Clades are defined by major variations of the SARS-CoV-2 genome, and this variant has been named GISAID Clade GV. To assess the epidemiological characteristics of the Clade GV, we conducted an ecological study[8] comparing the frequency in the Spanish Communities (Fig. 1b)[9] with epidemiological parameters including SARS-CoV-2 prevalence (number of cases per 1,000,000 people), CFR, and mortality (number of deaths per 1,000,000 people). Prevalence was positively correlated with Clade GV frequency (Fig. 1c), suggesting that the subtype is more transmissible than previously prevalent SARS-CoV-2. CFR tended to be lower in epidemic areas (Fig. 1c), suggesting that it is an attenuated virus in terms of lethality. Despite the lower CFR of the subtype, the very high prevalence led to increased mortality.

It is enigmatic why the COVID-19 mortality differs greatly between countries and regions. Since mortality is the product of prevalence and CFR, these factors were plotted in 67 countries (Fig. 2). We excluded countries with low PCR tests and uncertain statistics, as the analysis would be confusing if we included countries with inadequate population surveys. High mortality in European countries was primarily due to high CFR, and high mortality in American countries was primarily due to high prevalence. CFR was high in Western Europe,[11] Mexico, Canada, and Ecuador, where the prevalence was not

very high. In contrast, Chile, USA, Peru, and Brazil had high prevalence, but CFR was equal to or less than that of Asian countries such as Japan. Low prevalence was prominent in East and Southeast Asian countries, and low CFR was prominent in the Middle East and Singapore. These indicate that there are important regional factors that determine the spread and severity of COVID-19. In the epidemiological analysis of the nature of SARS-CoV-2, it was considered necessary to stratify by region in order to eliminate these factors.

Europe

We first focused on European countries and performed ecological study comparing SARS-CoV-2 genetic variations ("quasispecies") and epidemiological parameters (prevalence, CFR, and mortality) (Extended Data Table S1). GISAID Clade GR, GH, and G were predominant in Europe (Fig. 3a and 3b). As the Clade GV has expanded in Europe since the end of July changing the epidemiological aspect, we conducted an analysis as of 15 July 2020, before the Clade GV outbreak. Clade G was significantly positively correlated with high mortality and moderately positively correlated with both CFR and prevalence (Fig. 3c). Clade GR had a weak positive correlation with mortality (Spearman correlation coefficient $\rho = 0.29$, $P = 0.021$) and no significant correlation was observed with Clade GH ($\rho = 0.028$, $P = 0.83$). Clades G, GR, and GH share the Spike: D614G mutation, which increases viral infectivity.[6,7] Therefore, these results suggest that the N: RG203KR/ORF14: G50N mutation in Clade GR and the ORF3a: Q57H mutation in Clade GH reduced the virulence of the Spike: D614G variant. However, analysis of variants from these predominant Clades could not explain why CFR was high in Europe (Fig. 2).

Previous epidemiological analysis and viral interference with the influenza

epidemic curve revealed that three types of SARS-CoV-2, types S, K, and G, invaded Japan in sequence and established partial herd immunity.[5] As shown in the study, circumstantial evidence suggests that the type S and G correspond to ORF8: L84S (Clade S) and Spike: D614G variants, respectively. Calculations using epidemiological parameters and Fermi estimates led to the following equation:

$F = 4.7z - 166.20y + 175.58x$

where F is CFR and x, y, and z are exposure to type S, K, and G virus, respectively.[5] This formula predicts that the presence of type S virus will lead to high CFR. Therefore, we plotted the relationship between Clade S and CFR in Europe. As predicted by the formula, high CFR was observed in areas with a high ratio of Clade S (Fig. 4a), showing a significant positive correlation. This variant was also correlated with increased prevalence and mortality (Fig. 4a). Multiple regression analyses (MRA) comparing the involvement of Clade G, GR, GH, and S revealed that only Clade S contributed primarily to mortality (standardised partial regression coefficient = 0.452, $P = 0.0158$). Notably, except for Spain, Clade S accounts for less than 10% of the virus. This means that virus-intrinsic causes cannot explain such high CFR.

Spike: D614 is located in a region predicted to be an antibody epitope,[12] forming a partially exposed structure resembling an amphiphilic helix.[5] The D614G mutation is predicted to change an acidic residue to a hydrophobic one, causing structural changes that alter antigenicity (Fig. 4b), which converts the Clade S virus spike (D614) antibody to low affinity. This conversion to low-affinity spike antibody[13] is expected to induce antibody-dependent enhancement (ADE).[14]

United States

The epidemic in USA, unlike Europe, did not show the trends expected from immune responses to the SARS-CoV-2 virus.[5] Instead, significant positive correlation with the proportions of non-Hispanic black (NHB) was noted in prevalence, CFR, and mortality (Table 1, Extended Data Fig. 1a). Hispanic was positively correlated with prevalence and mortality (Table 1). Higher population density has been implicated to higher SARS-CoV-2 prevalence of NHB and Hispanic.[15] Indeed, population densities calculated based on the United States Census Bureau 2010 were positively correlated with CFR and mortality (Table 1, Extended Data Fig. 1b). However, MRA suggested that the proportion of NHB and Hispanic contributed more to the prevalence and mortality than population density (Table 2). In contrast, population density, rather than NHB and Hispanic race, seemed to be a major determinant of CFR.

Interestingly, the proportion of non-Hispanic American Indian was negatively correlated with CFR and mortality (Table 1, Extended Data Fig. 2). The states with Native American settlements may have low population densities. However, MRA suggested that non-Hispanic American Indian contributed significantly to CFR and mortality independent of the population density (Table 2). There are many national parks near Native American settlements, and people in such areas may have immunity that had been induced by exposure to microorganisms such as animal-derived viruses.[16] Indeed, the areas of national parks in the US states were negatively correlated with the CFR and mortality of SARS-CoV-2 (Table 1). However, MRA showed less contribution than Native Americans, suggesting that the presence of national parks alone cannot explain the decline in CFR and mortality. Overall, in USA, the living and social conditions indicated by population density and race may have had a significant impact on the epidemic outcome.

Latin America and Canada

Next, we analysed the effects of viral mutations in the countries of the Americas (Extended Data Table S2). Races was so influential in the US epidemic that USA was excluded to conduct a stratified analysis. In contrast to Europe, only Clade GR was positively correlated with mortality (Fig. 5a). Therefore, we investigated the possibility that new mutations that increase virulence were added to the Clade GR. The high prevalence in Peru was associated with the ORF1a: T1246I mutant originating from Clade GR probably in Croatia and spreading to Peru, South Africa, and New Zealand (Extended Data Fig. 3a). The ORF6: I33T variant was derived from Clade GR and spread within Brazil and to neighbouring countries including Chile (Extended Data Fig. 4a and 4c). Spike: V1176F variant also originated from Clade GR in Brazil, with the additional ORF1a: L3930F mutation associated with further spread (Extended Data Fig. 4b). N: S2F; ORF1a: T1250I; Spike: T307I variant originated from Clade GR in Chile with viral spread (Extended Data Fig. 4c). ORF1b: T2592I variant originated from Clade GR in Ecuador (Extended Data Fig. 3b). In Mexico, Clade GR has not generated new mutations leading to viral expansion (Extended Data Fig. 5a, right panel) with a low prevalence of SARS-CoV-2 (Extended Data Table S2). The emergence of new variants from Clade GR appears to have contributed to the high prevalence (Fig. 2) and mortality (Fig. 5a) in Peru, Chile, Brazil, and Ecuador (Extended Data Table S2).

CFR was higher in Mexico than other American countries (Extended Data Table S2). To assess which variants prevalent in Mexico boosted CFR, ecological study was conducted in Mexican states. N: S194L/ORF14: Q41* variant originated from Clade G probably in Finland and spread to Mexico, USA, Switzerland, Norway, and UK

(Extended Data Fig. 5a). MRA has shown that N: S194L/ORF14: Q41* was a major contributor to mortality in Mexican states (Extended Data Table S3). N: S194L/ORF14: Q41* tended to increase prevalence and CFR, thus increasing mortality (Extended Data Fig. 5b). The spread of this variant in USA (Extended Data Fig. 6) may have also contributed to the increased mortality in USA (Extended Data Table S2). Clade GH contributed to lower prevalence in Mexican states (Extended Data Table S3). The M: D3G variant originated from Clade G in Europe and spread to Peru and Ecuador (Extended Data Fig. 3a and 3b). ORF1a: S984G was derived from Clade G in Ecuador (Extended Data Fig. 3b, left panel). Since Clade G was associated with higher CFR in Europe (Fig. 3b), the N: S194L/ORF14: Q41*, M: D3G, and ORF1a: S984G variants from Clade G could account for high CFR in Mexico and Ecuador (Extended Data Table S2).

Panama had the second highest prevalence in the Americas after Chile (Extended Data Table S2). We investigated whether Panama had a special variant that boosted prevalence. N: S197L/ORF14: Q44*; ORF1a: F3071V; ORF3a: G196V variant was highly prevalent in Panama. This subtype originated from Clade S and spread from Panama to Spain, Australia, New Zealand, and Kazakhstan (Extended Data Fig. 7a). This subspecies appears to be responsible for high prevalence of SARS-CoV-2 in Panama.

Canada had the second highest CFR after Mexico (Extended Data Table S2), so we searched for mutations that boosted CFR in the country. ORF3a: T14I was epidemic in Quebec among Canadian provinces (Extended Data Fig. 7b, left panel). Ecological studies comparing Canadian provinces were conducted on variants including Spike: D614G, Clade GR, ORF1a: T265I, Clade GH Clade G, and ORF3a: T14I. MRA revealed a major contribution of the ORF3a: T14I variant to prevalence, CFR, and mortality (Extended Data Table S4). In addition, more mutations have occurred in the ORF1a:

T265I variant associated with virus spread (Extended Data Fig. 7b, middle panel). ORF3a: T14I and these mutants from ORF1a: T265I appear to be responsible for the high mortality in Canada. Branching of numerous mutants from ORF1a: T265I with viral spread has also been observed in USA (Extended Data Fig. 6) and may have also contributed to high prevalence in USA (Extended Data Table S2).

Then, we examined the countries with lower mortality in the Americas. A variant with ORF1b: A1844V mutation downstream of ORF1a: S3884L originated from ORF1a: T265I in Clade GH and spread to USA, Israel, Argentina, and Kazakhstan (Extended Data Fig. 8a). The N: S197L/ORF14: Q44* mutation was added in Argentina in connection with virus spread. Most of SARS-CoV-2 expanding in Suriname was a ORF1b: P909L variant from Clade G (Extended Data Fig. 8b). The spread of these variants did not result in increased mortality in epidemic countries. The absence of particularly widespread variants was characteristic of the epidemic in Costa Rica, Uruguay, and Jamaica, with low prevalence (Extended Data Table S2). Therefore, in the Americas, excluding USA, the presence or absence of the spread of new viral variants in that country seemed to have a significant impact on the epidemic outcome.

Africa

Among African countries with GISAID SARS-CoV-2 genome data (Extended Data Table S5), the proportion of Clade S showed a significant negative correlation with prevalence and mortality (Fig. 5b). This is consistent with the view that this S type is an attenuated virus.[17] Spike: D614G correlated with increased mortality (Fig. 5b), consistent with increased viral infectivity.[6,7]

In the early epidemics in China, the S type was the main type in areas other than

Wuhan.[17] Due to the large number of Chinese workers working in Africa, the virus may have been introduced early. Therefore, we calculated the ratio of Chinese migrant workers in 2018 to each country's population. MRA was performed including virus subtypes and Chinese worker ratio. Interestingly, the Chinese worker ratio contributed more strongly to prevalence and mortality than the Clade S (Table 3). Although the virus subtypes contributed significantly to CFR, there was still an independent contribution from the Chinese worker ratio. This result is consistent with the view that the early introduction of the attenuated virus by Chinese workers immunized Africans and prevented subsequent serious outbreaks.[18]

South Africa had the highest prevalence among African countries (Extended Data Table S5). ORF1a: Y4080H originated from Clade GR in South Africa and was associated with virus spread (Extended Data Fig. 9a). The ORF1a: T1246I from Clade GR, described above, also spread to South Africa. The ORF1b: P970L variant from Clade G was also widespread in South Africa (Extended Data Fig. 9a). The ORF1a: S3099L variant from Clade GR spread in The Gambia (Extended Data Fig. 9b). The emergence of ORF1a: Y4080H, ORF1a: T1246I, and ORF1a: S3099L mutants from Clade GR may be the reason why Clade GR contributed to increased prevalence and mortality in the African region (Table 3). The N: S187L/ORF14: H34Y mutation originated from Clade G and spread to Senegal and The Gambia (Extended Data Fig. 9b). Egypt's prevalence was low, but high CFR increased mortality, which is a pattern like Europe.

Moreover, the South African epidemic had a racial element. Prevalence in South African provinces was positively correlated with the proportion of white and coloured races (Fig. 5c). Since whites tend to live in densely populated areas, we conducted MRA of race and population density. Whites, not population densities, were the main

contributors to prevalence (Table 4). In contrast, coloured races and population densities contributed significantly to mortality (Table 4). These results suggest that the infection is widespread among whites in South Africa and that coloured races and overcrowded urban environment are involved in increasing mortality. Taken together, on the African continent except South Africa, the establishment of partial herd immunity to Clade S (Fig. 5b), probably through contact with Chinese migrant workers (Table 3), had a major impact on epidemic dynamics, with viral subtypes having subtle effects.

India

As of November 2020, India has the second highest number of infections and the third highest number of deaths in the world. We investigated the genotypes of viruses that are prevalent in India. The N: S194L/ORF14: Q41* variant from Clade GH (Q type, see below) spread to Gujarat, Delhi, and West Bengal. This variant was positively correlated with CFR in Indian states and Union Territory (Fig. 5d). MRA also showed that this variant contributed significantly to CFR (Table 5). The ORF1a: A1812D variant was derived from Clade GR, spread to Maharashtra, Gujarat, and Telangana (Extended Data Fig. 10a), and appears to be the main cause of the current epidemic in South India. MRA showed that ORF1a: A1812D contributed to prevalence and mortality (Table 5), but the variant did not tend to increase CFR (Extended Data Fig. 10b). In South India, the prevalence was reported to be low in the elderly, while the infection spread among children.[19] This pattern is known to occur in the elderly when a similar infection spread in their early childhood, as exemplified by the Spanish flu.[20] The presence of such cross-reactive immunity in the elderly may have reduced CFR from ORF1a: A1812D.[21] The ORF3a: L46F, derived from Clade GR and spreading within Telangana, tended to lower

CFR (Extended Data Fig. 10c) and contributed to lower mortality (Table 5). The lack of such a CFR reduction variant may be responsible for the high CFR in Tamil Nadu. The ORF1a: L3606F variant that originated from Clade O (tentatively named F type) (Extended Data Fig. 10a, right panel) contributed to prevalence (Table 5) but did not contribute to mortality, probably because it tends to reduce CFR. It is also noteworthy that population contributed to both prevalence and mortality (Table 5). Therefore, in India, population density and the characteristics of the spreading virus appear to have influenced epidemic outcomes.

Asia and Oceania

In this region, the N: S194L/ORF14: Q41* variant was derived from Clade GH and was widespread in Saudi Arabia, India, Australia, and New Zealand (Fig. 6a). This mutant, tentatively named Q type, was positively correlated with prevalence and mortality (Fig. 6b). Consistent with ecological studies in Indian states (Fig. 5d), there was a tendency to increase CFR (Fig. 6b). MRA showed that Q type was the major contributor to mortality (Table 6). In contrast, ORF1a: L3606F (F type) and Spike: D614G variants contributed to lower CFR. The F type occurred in the early stages of the epidemic and spread worldwide (Extended Data Fig. 11a). It is especially widespread in Singapore, Australia, and New Zealand. This variant correlated with a decrease in CFR (Fig. 6c).

ORF1a: G3334S originated from Clade GR and spread to Peru, Oman, and Serbia (Extended Data Fig. 11b). This subtype accounts for 75% of the Omani virus and may have contributed to the high prevalence in the country (Fig. 2; Extended Data Table S6). The ORF1b: A1844V; ORF1a: S3884L variant from ORF1a: T265I in Clade GH spread in Israel (Extended Data Fig. 8a and 12a). Another ORF1a: L3606F mutation

independently originated from ORF1a: T265I and spread to USA and Israel (Extended Data Fig. 12b). The ORF1a: L3606F mutation is frequently scattered in the phylogenetic tree of the virus (Extended Data Fig. 12c), suggesting a convergent evolution in favour of viral transmission. The N: S194L/ORF14: Q41* mutation that occurred after ORF1a: L3606F was also associated with spread of infection (Extended Data Fig. 12b). Like USA and Canada (Extended Data Fig. 6 and 7b), various mutations occurring downstream of the ORF1a: T265I subtype may have contributed to the increased prevalence in Israel (Extended Data Table S6). About 60% of Bahrain's infections are F type, which was associated with a high prevalence in the country (Extended Data Table S6).

Indonesia had the highest mortality rate in the East and Southeast Asian region due to its high CFR (Extended Data Table S6). An ecological study of the Indonesian provinces was conducted to identify the cause of high CFR. ORF1b: P218L originated from Clade GH and spread to East Java and Central Java (Extended Data Fig. 13a). ORF1a: D1532G was derived from Clade L and the spread of the virus was limited to Papua Province (Extended Data Fig. 13b). According to the MRA, ORF1b: P218L contributed to the increase in CFR and ORF1a: D1532G contributed to the decrease in CFR (Extended Data Table S7). Therefore, the relatively high CFR may be because of new mutations occurring in Indonesia.

Mortality rates in Australia and Bangladesh were comparable to Indonesia (Extended Data Table S6). ORF1a: I300F was derived from Clade GR and spread from Bangladesh to Australia in connection with viral expansion. In Bangladesh, another ORF1a: I300F mutation occurred independently from the Clade GR and was associated with virus spread (extended data Figure 14b, lower branch), suggesting convergent evolution. Overall, the ORF1a: I300F mutation appears to be involved in increased

prevalence in Bangladesh (extended data Figure 14c).

In Australia, Spike: S477N mutation occurred downstream of ORF1a: I300F, causing an infectious outburst confined to Australia (Extended Data Fig. 14a and 15a). An ecological study of Australian states was conducted to estimate the nature of this variant (tentatively named N type). MRA revealed that N type contributed to increased prevalence and mortality (Extended Data Table S8). N type tended to increase CFR (Extended Data Fig. 15b). The Spike: S477N mutation also occurred independently from Clade GH in Europe together with N: M234I, A376T; ORF1a: M3087I; ORF1b: A176S, V767L, K1141R, E1184D mutations (tentatively named DN type) and spread to France, Slovakia, and Norway (Extended Data Fig. 16a), suggesting that the Spike: S477N mutation is a convergent evolution that confer a selective advantage to the virus. Clade GV, DN type, and the ORF1a: H1113Y variant that originated from Clade G in Belgium (Extended Data Fig. 16b) seem to be the three major variants in the European epidemic as of November 2020.

Hong Kong had viral spread associated with Spike: S12F and N: A12G/ORF9b: H9D mutations in Clade GR (Extended Data Fig. 17a). The ORF1a: S3884L mutation was derived from ORF1a: T265I in Clade GH and spread to USA, Israel, Kazakhstan, Argentina, and South Korea (Extended Data Fig. 16b). Outbreaks in South Korea were associated with the ORF1b: Q2403L mutation downstream of ORF1a: S3884L (Extended Data Fig. 17b). Viral spread in Singapore was associated with the domestically prevalent ORF1a: S2015R mutation in F type (Extended Data Fig. 18a) and ORF1a: D3042N mutant from Clade GR (Extended Data Fig. 18b). Clade GV invaded Hong Kong and Singapore but did not spread. Partial herd immunity established in East and Southeast Asia may be able to suppress the spread of Clade GV. Alternatively, already prevalent

virus subtypes may be competing with the Clade GV. Populated areas such as Hong Kong and Singapore may increase selective pressure due to virus competition. In contrast to Clade GV, Q type is beginning to spread in Singapore (Fig. 6a, right panel; Extended Data Fig. 18a, right panel) and Hong Kong (Extended Data Fig. 17a, right panel; Extended Data Fig. 18c), demonstrating that Q type can surpass either regional immunity or virus competition. In summary, there are differences in the degree of herd immunity between the eastern and western parts of the Asia-Oceania region, and against this background, the spread of different viral subtypes in each country may have determined the epidemic outcome.

Time, Place, Mutation

To infer the molecular mechanism of how viral protein mutations alter pathological processes, we created a "spot map" of the location of mutations on the SARS-CoV-2 genome (Fig. 7a). Open reading frame (ORF) 1a and 1b encode non-structural proteins (nsp) that compose the viral replication and transcription complex.[22] ORF3a to 14 encode accessory proteins that are not required for intracellular viral replication but are thought to play roles in the natural host. Mutations associated with viral expansion were concentrated in proteins involved in viral growth. Analyses of SARS-CoV-2 and other coronaviruses have revealed the key functions of the viral proteins.[23,24] Nsp3, nsp4, and nsp6 induce double membrane vesicles of replication organelles.[22] Nsp2 is localized in the replication complex with nsp3. Nsp12 is RNA-dependent RNA polymerase. Nsp7 and nsp8 are cofactors of nsp12. ORF6 interacts with nsp8 and enhances viral replication *in vitro* and *in vivo*. Nsp5 is the main protease that cleaves viral polyprotein and yields nsps. Nsp13 is a helicase and RNA 5'-triphosphatase.

Nsp14 is involved in proofreading. Nsp14 and nsp16 mediate RNA cap formation. The Spike: S477N mutation has been shown to enhance folding and binding to the angiotensin-converting enzyme 2 (ACE2) receptor.[25,26] Consistent with our results, Spike: S477N has been correlated with higher CFR.[25] Spike: V1176F mutation has been predicted to increase stability of the protein and flexibility of the trimeric stalk.[25] It is also reported that the Spike: V1176F, ORF6: I33T, and ORF1a: L3930F mutations are correlated with increased CFR in Brazil.[25] Overall, the mutations may improve the fitness of SARS-CoV-2 in the cell and promote the spread of the virus.

To understand the trends of SARS-CoV-2 mutations involved in virus spread, we potted an "epidemic curve" of the mutations. Virus spread mutations peaked in March 2020 and have since peaked out (Fig. 7b), suggesting that some mechanisms that peak the mutations has worked and the mutation epidemic is ending. However, it is crucial to continue monitoring whether new mutations related to virus spread will turn to an increasing trend.

Discussion

A global comparison of SARS-CoV-2 prevalence and CFR shows that high mortality is due to high CFR in Europe and high prevalence in Latin America. In high-prevalence countries such as Latin America, new mutations associated with the spread of the virus have emerged. New mutations can be both the cause and consequence of a viral epidemic. Alternatively, it may be a founder effect due to the accidental spread of the mutated virus. However, the fact that there are new variants in high-prevalence countries on the same continent and not in low-prevalence countries contradicts the founder effect. Convergent evolution has been suggested for many mutations such as ORF1a: I300F (nsp2), ORF1a: A1812D (nsp3), ORF1a: L3606F (nsp6), Spike: S477N, Spike: D614G, M: D3G, ORF8: L84S, N: S194L/ORF14: Q41*, and N: S197L/ORF14: Q44*, which may improve viral fitness. Mutants did not develop gradually as the virus spread, appeared primarily at the beginning of the epidemic, suggesting that the mutants are more likely to be the cause than the result of virus spread. A global analysis of recurrent mutations in SARS-CoV-2 reported that the mutations did not significantly increase the number of descendants.[27] However, as the present study has revealed, transmissible SARS-CoV-2 variants expand in niches that reflect population heterogeneity and avoid defensive herd immunity and interference with other subtypes. Even if the subtypes are highly transmissible, they cannot spread freely throughout the world, so a collective analysis of the world's subtypes[27] does not reveal a niche spread. Such context-sensitive propagation of the SARS-CoV-2 subtype could only be visualized by scrutinizing the state of local spread.

The impact of the D614G mutation on mortality varied from region to region,

suggesting that this mutation affects the pandemic by multiple mechanisms. The antiviral immune responses play a central role in viral cell damage, and the SARS-CoV-2 appears to be no exception.[28,29] Thus, not only the effects of viral mutations on infected cells, but also the immune responses to the mutant virus are likely to affect the pathogenesis. The adaptive immune response is strongly influenced by previous viral infections that induce allogeneic or cross-immunity.[21] Therefore, we hypothesized that the major factor diversifying the effects of D614G mutation in different parts of the world might be the immunity acquired against the previously prevalent viruses. In Europe , mortality due to the D614G variant has increased significantly in countries with Clade S. This illustrates that immunity to Clade S altered the immune responses to the D614G virus and dramatically increased lethality.

Spike: D614G mutation caused a major structural change in the epitope of the antibody, suggesting that the antibody against the D614 spike becomes a low affinity antibody for the G614 spike. It is estimated that the G614 virus in Europe caused ADE by the low-affinity antibody in areas where the Clade S virus had been present, leading to increased CFR and mortality.[5] Alternatively, antibodies with poor neutralizing activity can form immune complexes to activate complement, causing conditions like vaccine-associated respiratory disease (VAERD).[13,30] Heterologous serum from D614 infected hamster can neutralize G614 virus efficiently,[6,7] but further analysis using monoclonal antibodies would be necessary. Furthermore, the definitive clinical effect of convalescent plasma has not been demonstrated[31] and it is desirable to investigate the effect of D614 antibodies *in vivo* in animals. Even neutralizing antibodies may induce antibody-dependent cellular phagocytosis (ADCP)[32] and release cytokines from cells with Fc receptors, leading to disease exacerbation. Indeed, it has been reported that a patient

reinfected with SARS-CoV-2 had more severe COVID-19 than the initial infection.[33] Without any other rational explanations, the results support our hypothesis that immunity to type S virus underlies the exacerbation of COVID-19 in Europe.[5]

Serosurveys in Kenya have suggested that the epidemic in Africa was like that in the West;[34] however, in contrast to the West, most patients were asymptomatic or mildly ill. On the African continent, in contrast to Europe, many of the epidemics were due to Clade S. Our previous analysis suggested that the outbreak of type K virus after type S offsets the effect of type S to increase CFR.[5] As a result, Clade S may have reduced prevalence in Africa without increasing CFR. Epidemiological analyses in the present study demonstrated that the attenuated SARS-CoV-2 epidemic on the African continent[18] was due to the influx of attenuated viruses by Chinese migrant workers.

Epidemiological effects provide further evidence that viral protein mutations alter the host's immune responses. Mutations associated with high CFR were concentrated in proteins involved in the regulation of innate immunity such as nsp3 and ORF3a. Nsp3 attenuates type I interferon responses.[35] ORF3a is an ion channel that activates NLRP3 inflammasome. A peptide spanning ORF3a: T14I has been reported as an HLA-DR T cell epitope,[36] potentially affecting cell-mediated immunity. The ORF3a: T14I mutation is also present in peptides predicted as HLA Class I T cell epitopes.[12] M: D3G and N: S194L mutations are present in antibody epitopes,[16] which may reduce antibody avidity. Mutation effects can contribute to immunomodulation that increases CFR, because innate immune system depression,[37] adaptive immune system dysregulation,[38] and incoordination between cell-mediated and humoral immunity[39] underlie the exacerbation of COVID-19.

Interestingly, we also found that mutations associated with higher prevalence,

such as ORF1a: I300F (nsp2), ORF1a: H1113Y (nsp3), ORF1b: A1844V (nsp14), S: T307I; ORF6: I33T, N: A12G/ORF9b: H9D, N: S197L, and N: A220V, are present in antibody epitopes.[16] ORF6: I33T is also located in a HLA-DR T cell epitope.[36] S: T307I and N: A220V are present in peptides predicted as HLA-DR and HLA Class I T cell epitopes, respectively.[12] Disruption of humoral and cell-mediated immune systems by viral mutations may also enhance the transmission of SARS-CoV-2 and increase prevalence. Historically, the most serious effect of mutations in pandemic viruses has been to avoid immune responses. Indeed, ORF1a: I300F (nsp2), M: D3G, and N: S194L/ORF14: Q41* are recurrent mutations in the phylogenetic tree, suggesting that immune evasion is a part of convergent evolution. Changes in immunity caused by the virus, as well as the immune status obtained from previous infections, appear to determine the pathogenicity of SARS-CoV-2.

High population densities not only make social distance difficult in urban areas but are also an indicator of poverty. Population density contributed to prevalence and mortality in India, while it was a determinant of CFR in USA. It should be noted that this is a problem that cannot be solved by usual Infectious disease control measures such stay-at-home order (lockdown) or quarantine of infected persons.

In USA and South Africa, known as melting pots and salad bowls, racial factors have contributed significantly to the spread of SARS-CoV-2. There is no evidence of exacerbation of COVID-19 in NHB,[40] arguing against racial biological vulnerabilities. The high prevalence among NHB and Hispanics has been attributed to living in densely populated areas.[15] However, our ecological studies suggest that prevalence and mortality are more racially influenced than population density. The spread of SARS-CoV-2 may have been due to various additional sociodemographic factors, such as a multi-

generational lifestyle, poor access to health care, inadequate hygiene measures due to poverty, crowded religious organizations and grocery stores,[41] and inability to reduce mobility[41] due to occupations that defy the shift to remote work. It is unclear why CFR and mortality were lower in states with higher proportions of non-Hispanic American Indians, but it may be due to interference from other infections or to lifestyle differences. In South Africa, whites were a determinant of morbidity stronger than population density, suggesting that infection is widespread in poor white settlements where social distancing and frequent hand washing are not feasible.[42] It is a historical lesson that the rich and the poor are equally affected when the masses remain unhealthy. Epidemic in these countries may not cease without improving racial inequality and ethnic disparity.

In the winter epidemic of 2020, the effects of cold weather will be inevitable in the Northern Hemisphere. Lower temperature and humidity weaken intrinsic, innate, and adaptive immune responses in the respiratory mucosa, facilitating and aggravating respiratory virus infections.[43] SARS-CoV-2 infection, which were often asymptomatic during the summer, will also become apparent, and the increase in critically ill patients will put substantial pressure on the healthcare system. The second wave of epidemics in Europe is due not only to seasonal factors, but also to the emerging subtypes such as Clade GV. In contrast to our results, Clade GV has been reported to increase CFR.[25] However, because Clade GV is widespread in Europe, where CFR is high, regional differences become confounding factors and apparently increase the Clade's CFR. Our ecological analysis conducted exclusively in Spain to eliminate confounding factors is expected to more accurately reflect the nature of the virus. Clade GV failed to spread in Singapore and Hong Kong, suggesting that the established herd immunity in Southeast and East Asia can suppressed this subtype. CD4 T lymphocytes against other common

cold coronaviruses are known to cross-react with SARS-CoV-2.[21,44] It is likely that the regions of Southeast Asia and East Asia around China have long been exposed to Chinese bat-derived coronaviruses[45] and have accumulated a large number of memory CD4 T lymphocytes, explaining the low prevalence and CFR of SARS-CoV-2 in this area. In contrast to other regions, the D614G virus tends to lower CFR in Asia and Oceania. In areas where adequate protective immunity has been established, COVID-19 by G614 virus may become less severe because of the tendency of the G614 virus to be neutralized by antibodies.[7]

The N type virus seems to be prevalent only in Australia so far. This epidemic pattern may be due to the Southern Hemisphere being in the winter season. Countries in the Northern Hemisphere need to be prepared for the possibility that this highly transmissible and virulent virus invades and spread in winter. Another dangerous virus is the Q type. Migrant workers working in the Middle East may have brought it home. The virus has increased CFR in India and increased mortality in the Middle East, South Asia and Oceania. In contrast to the Clade GV, the Q type has spread in Singapore and Hong Kong. Q type that originated in the Middle East may have undergone a kind of "immunoediting" in the local immune environment. People in East Asia may be vulnerable to Middle Eastern coronaviruses such as MERS, as illustrated by the South Korean pandemic.[46] The inability of Southeast and East Asian immunity to control the spread of this virulent subtype poses a threat of spread and death in these regions.

Taken together, the above considerations indicate that mutations drive the SARS-CoV-2 epidemic. History has shown that if an infectious disease with a basic reproduction number (R_0) ~2 freely spreads in the population, it will reach herd immunity in about 3 weeks,[20] as was the case in Wuhan, China.[47] Highly transmissible subtypes

have a higher R_0 and thus raise the herd immunity threshold. Therefore, even if herd immunity is reached, when a virus subtype with a higher R_0 invades, additional infected persons will have to come out. In addition, if a viral mutation that evades immunity occurs, the existing herd immunity will no longer work, and the epidemic will resume. The protracted SARS-CoV-2 pandemic may be due to the repeated virus mutations overtaking herd immunity. This indicates that the infectious epidemic will not cease unless the mutation epidemic ends. Gene mutations continue to occur in the process of transmission from person to person because of the error-prone viral FNA polymerase and RNA editing.[27] However, even if mutations continue to occur, they are not selected unless they are new protein changes that further increase the efficiency of virus transmission. As the mutation epidemic curve shows, the emergence of new mutations associated with virus spread have peaked out, strongly suggesting that there is little room for mutations to make more fit viruses. If no new mutations appear, the epidemic will end when herd immunity against the current virus is established. We predict that the present SARS-CoV-2 epidemic will end by March 2021 when winter ends in the Northern Hemisphere and mucosal defences and immune responses to winter respiratory viruses are restored.[43]

This study also suggested the possibility of competition between SARS-CoV-2 subtypes. Viruses with high CFR are difficult to spread among humans, and there are many effective countermeasures to prevent their spread. Thus, the ultimate winner of virus competition is the virus with low CFR and high infectivity. In Hong Kong and Singapore, viral competition has been suggested to prevent the transmission of newly invaded variants by existing highly contagious variants. This means that the “mild cold” subtypes are beginning to drive out the “severe pneumonia” subtypes. Immunity to common cold coronaviruses lasts only about 10 months.[2,30] Even if the epidemic ends in

spring 2021 in the Northern Hemisphere, the immune system will weaken in autumn and the SARS-CoV-2 epidemic will resume. However, as long as the current trend of attenuation continues, SARS-CoV-2 is likely to return as a mere cold virus.[2,3]

Our study has some limitations. First, there are unavoidable problems inherent to epidemiological studies. Statistical analysis yields correlations, not necessarily causal relationships. It takes time and many experimental approaches to reach a true conclusion about causality. Second, the virus genome data of each country registered in GISAID do not always accurately reflect the virus prevalent in that country. Furthermore, there are many areas where genomic data are not registered in GISAID. Deviations from the population are inherent in random sampling, and we must strive to make estimates that reflect reality as much as possible from existing data. Third, there are limits to human information processing. We manually processed and interpreted a huge amount of genomic data, obtained epidemiological information, and organized statistical results, but in order to be more reliable, processing using a supercomputer or AI will be necessary. However, since it is important to have a bird's-eye view, we believe that human processing processes are also important, not just relying on machines. Finally, the theoretical framework of research has not been established. The science of infection pandemics requires integration as a comprehensive discipline that includes medicine, biology, microbiology, cultural anthropology, sociology, political science, and economics. Experts in individual fields have accumulated experience and continued trial and error. However, considering the disagreements and social divisions in this pandemic countermeasure, it is imperative to build an academic system on a secure foundation.

This epidemiological study revealed the nature of epidemic SARS-CoV-2 virus by subtyping the virus using “genetic fingerprinting”[8] and comparing the subtypes with

epidemiological information. We also outlined the pandemic using other indicators that may affect the transmission and severity of infections. Although the overview is complex, several factors have been identified that determine pandemic dynamics, including viral genotype, regional herd immunity, population density, and race. Parallel research that addresses individual problems found in this comprehensive analysis will help to uncover the whole picture of the world suffering from the pandemic and create effective means of ending it.

Methods

Sources of data

Data were obtained from websites including the Worldometer,[48] Wikipedia,[49-53] Governing,[54] State & County Rankings,[55] China Africa Research Initiative,[56] [World economy newsletter],[57] and the Ministry of Health and Family Welfare, Government of India.[58] The GISAID database[9] was used for the phylogenetic analyses of SARS-CoV-2 gene mutations. The country and race names used on these websites were adopted.

Modelling analysis

Mathematical modelling was performed according to the practice of theoretical epidemiology.[8] Helix-wheel projections were created to analyse amphiphilic helix structures using the NetWheels projections maker. Statistical analyses were performed with the use of the Statcel4 add-in package (OMS Publishing, Tokorozawa, Japan) for Microsoft Excel.

Data availability

All data are available on request.

References

1 Hu, B., Guo, H., Zhou, P. & Shi, Z. L. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol*, doi:10.1038/s41579-020-00459-7 (2020).

2 Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**, 860-868, doi:10.1126/science.abb5793 (2020).

3 Saad-Roy, C. M. *et al.* Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. *Science* **370**, 811-818, doi:10.1126/science.abd7343 (2020).

4 Kamikubo, Y. & Takahashi, A. Epidemiological Tools that Predict Partial Herd Immunity to SARS Coronavirus 2. *medRxiv*, 2020.2003.2025.20043679, doi:10.1101/2020.03.25.20043679 (2020).

5 Kamikubo, Y., Hattori, T. & Takahashi, A. Paradoxical dynamics of SARS-CoV-2 by herd immunity and antibody-dependent enhancement. *Cambridge Open Engage.*, doi:10.33774/coe-2020-fsnb3-v2 (2020).

6 Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science*, doi:10.1126/science.abe8499 (2020).

7 Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, doi:10.1038/s41586-020-2895-3 (2020).

8 Giesecke, J. *Modern Infectious Disease Epidemiology*. 3rd edn, (CRC Press, 2017).

9 GISAID. *Genetic epidemiology of hCoV-19*, <https://www.gisaid.org/epiflu-applications/phylodynamics/> (2020).

10 Hodcroft, E. B. *et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*, 2020.2010.2025.20219063, doi:10.1101/2020.10.25.20219063 (2020).

11 O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, doi:10.1038/s41586-020-2918-0 (2020).

12 Grifoni, A. *et al.* A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* **27**, 671-680 e672, doi:10.1016/j.chom.2020.03.002 (2020).

13 Bournazos, S., Gupta, A. & Ravetch, J. V. The role of IgG Fc receptors in antibody-dependent enhancement. *Nat Rev Immunol* **20**, 633-643, doi:10.1038/s41577-020-00410-0 (2020).

14 Eroshenko, N. *et al.* Implications of antibody-dependent enhancement of infection for SARS-CoV-2 countermeasures. *Nat Biotechnol* **38**, 789-791,

doi:10.1038/s41587-020-0577-1 (2020).

15 Vahidy, F. S. *et al.* Racial and ethnic disparities in SARS-CoV-2 pandemic: analysis of a COVID-19 observational registry for a diverse US metropolitan population. *BMJ Open* **10**, e039849, doi:10.1136/bmjopen-2020-039849 (2020).

16 Shrock, E. *et al.* Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science* **370**, doi:10.1126/science.abd4250 (2020).

17 Lu, J. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review* **7**, 1012-1023, doi:10.1093/nsr/nwaa036 (2020).

18 Mbow, M. *et al.* COVID-19 in Africa: Dampening the storm? *Science* **369**, 624-626, doi:10.1126/science.abd3902 (2020).

19 Laxminarayan, R. *et al.* Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science* **370**, 691-697, doi:10.1126/science.abd7672 (2020).

20 Crosby, A. W. *America's Forgotten Pandemic: The Influenza of 1918*. (Cambridge University Press, 2003).

21 Sette, A. & Crotty, S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. *Nat Rev Immunol* **20**, 457-458, doi:10.1038/s41577-020-0389-z (2020).

22 V'Kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol*, doi:10.1038/s41579-020-00468-6 (2020).

23 Liu, D. X., Fung, T. S., Chong, K. K., Shukla, A. & Hilgenfeld, R. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res* **109**, 97-109, doi:10.1016/j.antiviral.2014.06.013 (2014).

24 Yoshimoto, F. K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. *Protein J* **39**, 198-216, doi:10.1007/s10930-020-09901-4 (2020).

25 Farkas, C., Mella, A. & Haigh, J. J. Large-scale population analysis of SARS-CoV-2 whole genome sequences reveals host-mediated viral evolution with emergence of mutations in the viral Spike protein associated with elevated mortality rates. *medRxiv*, 2020.2010.2023.20218511, doi:10.1101/2020.10.23.20218511 (2020).

26 Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310 e1220, doi:10.1016/j.cell.2020.08.012 (2020).

27 van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent

mutations in SARS-CoV-2. *Nat Commun* **11**, 5986, doi:10.1038/s41467-020-19818-2 (2020).

28 Tay, M. Z., Poh, C. M., Renia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* **20**, 363-374, doi:10.1038/s41577-020-0311-8 (2020).

29 Vabret, N. *et al.* Immunology of COVID-19: Current State of the Science. *Immunity* **52**, 910-941, doi:10.1016/j.immuni.2020.05.002 (2020).

30 Sariol, A. & Perlman, S. Lessons for COVID-19 Immunity from Other Coronavirus Infections. *Immunity* **53**, 248-263, doi:10.1016/j.immuni.2020.07.005 (2020).

31 Simonovich, V. A. *et al.* A Randomized Trial of Convalescent Plasma in Covid-19 Severe Pneumonia. *N Engl J Med*, doi:10.1056/NEJMoa2031304 (2020).

32 Pierce, C. A. *et al.* Immune responses to SARS-CoV-2 infection in hospitalized pediatric and adult patients. *Sci Transl Med* **12**, doi:10.1126/scitranslmed.abd5487 (2020).

33 Tillett, R. L. *et al.* Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect Dis*, doi:10.1016/S1473-3099(20)30764-7 (2020).

34 Uyoga, S. *et al.* Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors. *Science*, doi:10.1126/science.abe1916 (2020).

35 Shin, D. *et al.* Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature* **587**, 657-662, doi:10.1038/s41586-020-2601-5 (2020).

36 Nelde, A. *et al.* SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat Immunol*, doi:10.1038/s41590-020-00808-x (2020).

37 Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718-724, doi:10.1126/science.abc6027 (2020).

38 Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463-469, doi:10.1038/s41586-020-2588-y (2020).

39 Rydyznski Moderbacher, C. *et al.* Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. *Cell* **183**, 996-1012 e1019, doi:10.1016/j.cell.2020.09.038 (2020).

40 Yehia, B. R. *et al.* Association of Race With Mortality Among Patients Hospitalized With Coronavirus Disease 2019 (COVID-19) at 92 US Hospitals. *JAMA Netw Open* **3**, e2018039, doi:10.1001/jamanetworkopen.2020.18039 (2020).

41 Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, doi:10.1038/s41586-020-2923-3 (2020).
42 Nordling, L. 'Our epidemic could exceed a million cases' — South Africa's top coronavirus adviser. *nature* **583**, 672 (2020).
43 Moriyama, M., Hugentobler, W. J. & Iwasaki, A. Seasonality of Respiratory Viral Infections. *Annu Rev Virol* **7**, 83-101, doi:10.1146/annurev-virology-012420-022445 (2020).
44 Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* **370**, 89-94, doi:10.1126/science.abd3871 (2020).
45 Latinne, A. *et al.* Origin and cross-species transmission of bat coronaviruses in China. *Nat Commun* **11**, 4235, doi:10.1038/s41467-020-17687-3 (2020).
46 Arabi, Y. M. *et al.* Middle East Respiratory Syndrome. *N Engl J Med* **376**, 584-594, doi:10.1056/NEJMsr1408795 (2017).
47 Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **382**, 1199-1207, doi:10.1056/NEJMoa2001316 (2020).
48 Worldometer. *COVID-19 Coronavirus Pandemic*, <https://www.worldometers.info/> (2020).
49 Wikipedia. *COVID-19 pandemic by country and territory*, <https://en.wikipedia.org/wiki/COVID-19_pandemic_by_country_and_territory> (2020).
50 Wikipedia. *List of national parks of the United States*, <https://en.wikipedia.org/wiki/List_of_national_parks_of_the_United_States> (2020).
51 Wikipedia. *Provinces of South Africa*, <https://en.wikipedia.org/wiki/Provinces_of_South_Africa> (2020).
52 Wikipedia. Demographics of South Africa. (2020).
53 Wikipedia. List of states and union territories of India by population. (2020).
54 Governing. *State Population By Race, Ethnicity Data*, <https://www.governing.com/gov-data/census/state-minority-population-data-estimates.html> (2020).
55 Rankings, S. a. C. *USA Population Density Ranking (By State)*, <http://us-ranking.jpn.org/PopulationDensity.html> (2020).
56 Initiative, C. A. R. *DATA: CHINESE WORKERS IN AFRICA*, <http://www.sais-cari.org/data-chinese-workers-in-africa> (2020).
57 Kawauchi, A. *[Africa Population Density Ranking]*,

<https://ecodb.net/ranking/area/G/imf_area_lp.html> (2020).
58 Government of India, M. *COVID-19 Statewise Status*, <https://www.mohfw.gov.in/> (2020).

Acknowledgements We thank N. Takebayashi for data on Chinese in Africa and India; Y. Ota, T. Sakamoto, R. Taniguchi for collecting data from GISAID. This work was supported by Grant-in-Aid for Scientific Research (KAKENHI; 17H03597, 16K14632, and JP17H01690) from the Japan Society for the Promotion of Science.

Author contributions YK and AT conceived the study and performed the analysis. AT collected data, performed the statistical analyses, and wrote the first draft of the manuscript. TH and AT analysed protein structures. YK, TH, and AT discussed the results and contributed to revisions of the manuscript.

Competing interests The authors declare no competing interests.

Correspondence and requests for materials should be addressed to Y.K. and A.T.

Table 1. Spearman correlation coefficient of ethnicity, population density, and national parks with prevalence, CFR, and mortality in US states.

	prevalence	CFR	mortality
non-Hispanic black (NHB)	0.624 ($P = 1.02 \times 10^{-5}$)	0.354 ($P = 0.0124$)	0.676 ($P = 1.74 \times 10^{-6}$)
Hispanic	0.297 ($P = 0.0358$)	0.0975 ($P = 0.491$)	0.330 ($P = 0.0195$)
non-Hispanic American Indian	-0.0831 ($P = 0.502$)	-0.532 ($P = 0.000103$)	-0.455 ($P = 0.000877$)
population density	0.225 ($P = 0.112$)	0.657 ($P = 3.36 \times 10^{-6}$)	0.673 ($P = 1.92 \times 10^{-6}$)
area of national park	-0.185 ($P = 0.0770$)	-0.295 ($P = 0.00967$)	-0.293 ($P = 0.0100$)

The prevalence, CFR, and mortality in US states as of 12 September 2020 were used for analyses.

Table 2. Multiple regression analyses of race and population density in the United States.

response variable	explanatory variable	standardised partial regression coefficient	*P* -value
prevalence	NHB	0.790	6.38×10^{-8}
	Hispanic	0.357	
	population density	-0.291	
CFR	population density	0.215	$P = 0.0278$
	non-Hispanic American Indian	-0.279	
mortality	NHB	0.380	$P = 0.00212$
	Hispanic	0.289	
	population density	0.0687	
	non-Hispanic American Indian	-0.188	

Table 3. Multiple regression analyses of African countries.

response variable	explanatory variable	standardised partial regression coefficient	*P* -value
prevalence	Clade GR	0.779	0.01158
	Chinese worker ratio	-0.592	
CFR	Spike: D614G	4.752	0.0355
	Clade S	4.400	
	Clade GH	-3.167	
	ORF1a: T265I	1.20	
	Clade G	-4.949	
	non-T265I	4.64	
	Chinese worker ratio	0.420	
mortality	Clade GR	0.762	0.0153
	Chinese worker ratio	-0.582	

Epidemiological and GISAID data as of 29 July 2020 were used for analysis. non-T265I, Spike: D614G variant except ORF1a: T265I.

Table 4. Multiple regression analyses of race and population density in South Africa.

response variable	explanatory variable	standardised partial regression coefficient	*P* -value
prevalence	White	0.761	0.0172
mortality	Coloured	4.20	0.00950
	Other	-5.89	
	African	-2.23	
	population density	1.16	

The prevalence, CFR, and mortality in South African provinces as of 17 September 2020 were used for analyses.

Table 5. Multiple regression analyses of Indian states and union territories.

response variable	explanatory variable	standardised partial regression coefficient	*P* -value
prevalence	ORF1a: A1812D	0.687	0.0302
	ORF1a: L3606F	0.356	
	population density	0.356	
CFR	Q type	0.676	0.0302
mortality	ORF1a: A1812D	0.712	0.000167
	population density	0.675	
	ORF3a: L46F	-0.356	

Epidemiological and GISAID data as of 23 September 2020 were used for analysis.

Table 6. Multiple regression analyses of Asia and Oceania.

response variable	explanatory variable	standardised partial regression coefficient	*P* -value
CFR	Spike: D614G	-0.668	0.000630
	ORF1a: L3606F	-0.842	
mortality	Q type	0.709	5.67×10^{-6}

Epidemiological and GISAID data as of 25 September 2020 were used for analysis.

Figure legends

Fig. 1 | The spread of SARS-CoV-2 Clade GV in Europe and the epidemic outcomes in Spain. a, Phylogenetic tree showing mutational branching of GISAID Clade GV (left panel) and geographical distribution of Clade GV among European countries (right panel) as of 11 November 2020. The color of the circle is different for each country. **b,** Pie charts showing the percentages of Clade GV in the Spanish Communities as of 24 August 2020. **c,** Correlation between the percentage of Clade GV and the prevalence, CFR, and mortality of COVID-19 in the Spanish Communities. ρ: Spearman correlation coefficient.

Fig. 2 | Prevalence and CFR of COVID-19 in major countries of the world. A scatter plot showing the relationship between prevalence (cases per 1,000,000) and CFR (%) in 67 countries as of 15 July 2020, excluding countries with few PCR tests (less than 1300 tests per 1,000,000 people) and countries with uncertain statistics. The countries in Europe, Americas, Africa, the Middle East, and Asia/Oceania are shown in green, magenta, light blue, purple, and brown, respectively.

Fig. 3 | Epidemic outcome in Europe dominated by GISAID Clade G. a, Epidemic of mainstream Clade GR, GH, and G in Europe as of 23 October 2020. The color of the circle is different for each Clade. **b,** Pie charts showing the ratio of Clades in European countries. **c,** Clade G was positively correlated with high prevalence, CFR, and mortality in Europe as of 15 July 2020. ρ: Spearman correlation coefficient.

Fig. 4 | Increased mortality associated with SARS-CoV-2 Clade S in Europe and the underlying molecular mechanism. a, Correlation between Clade S (%) and the

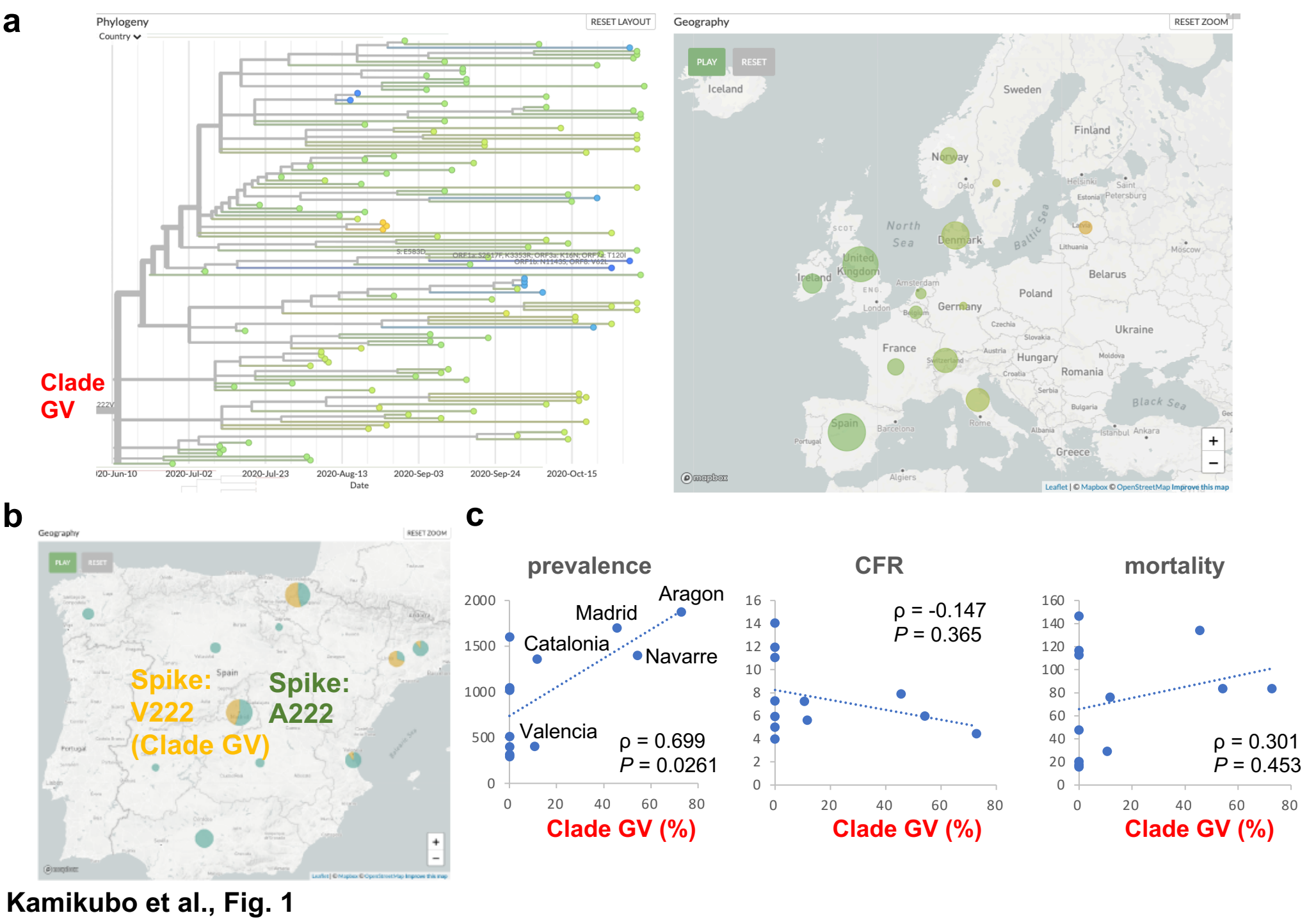
prevalence, CFR, and mortality in major European countries as of 15 July 2020. ρ: Spearman correlation coefficient. **b,** Impact of the Spike: D614G mutation on structures analysed by Helix-wheel projections.

Fig. 5 | SARS-CoV-2 variants and racial factors that determined epidemic outcomes in the Americas, Africa, and Indian Peninsula. a, Clade GR was positively correlated with mortality in the Americas as of 15 July 2020. **b,** Clade S showed a significant negative correlation with morbidity and mortality in African countries as of 29 July 2020. Spike: D614G was positively correlated with mortality. **c,** Prevalence in South African states was positively correlated with the proportion of Whites and Coloured as of 17 September 2020. **d,** Q type was positively correlated with CFR in Indian states and Union Territory as of 23 September 2020. ρ: Spearman correlation coefficient.

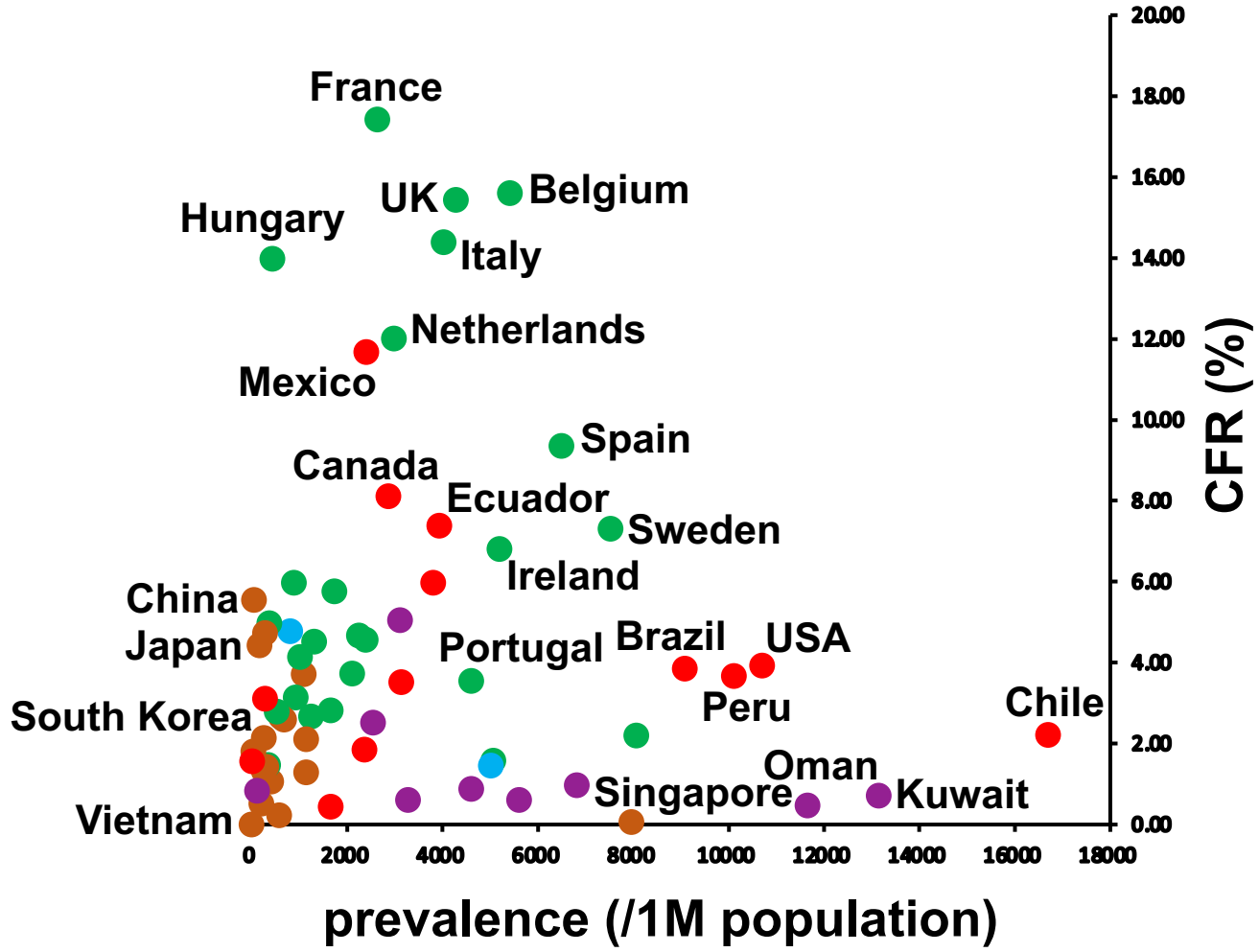
Fig. 6 | Q type in Asia and Oceania associated with Increased mortality. a, Mutant branching of Q type (N: S194L/ORF14: Q41* from Clade GH) (left panel) and dissemination to Saudi Arabia, India, Australia, and New Zealand (right panel) as of 18 November 2020. The color of the circle is different for each country. **b,** Q type showed positive correlation with prevalence, CFR, and mortality in Asian countries as of 25 September 2020. **c,** F-type was negatively correlated with mortality in Asian countries. ρ: Spearman correlation coefficient.

Fig. 7 | Epidemiology of SARS-CoV-2 mutations. a, "Spot map" of mutation locations on the SARS-CoV-2 genome. Amino acid changes caused by mutations associated with viral spread are shown above the genomic diagram. Mutations associated with increased

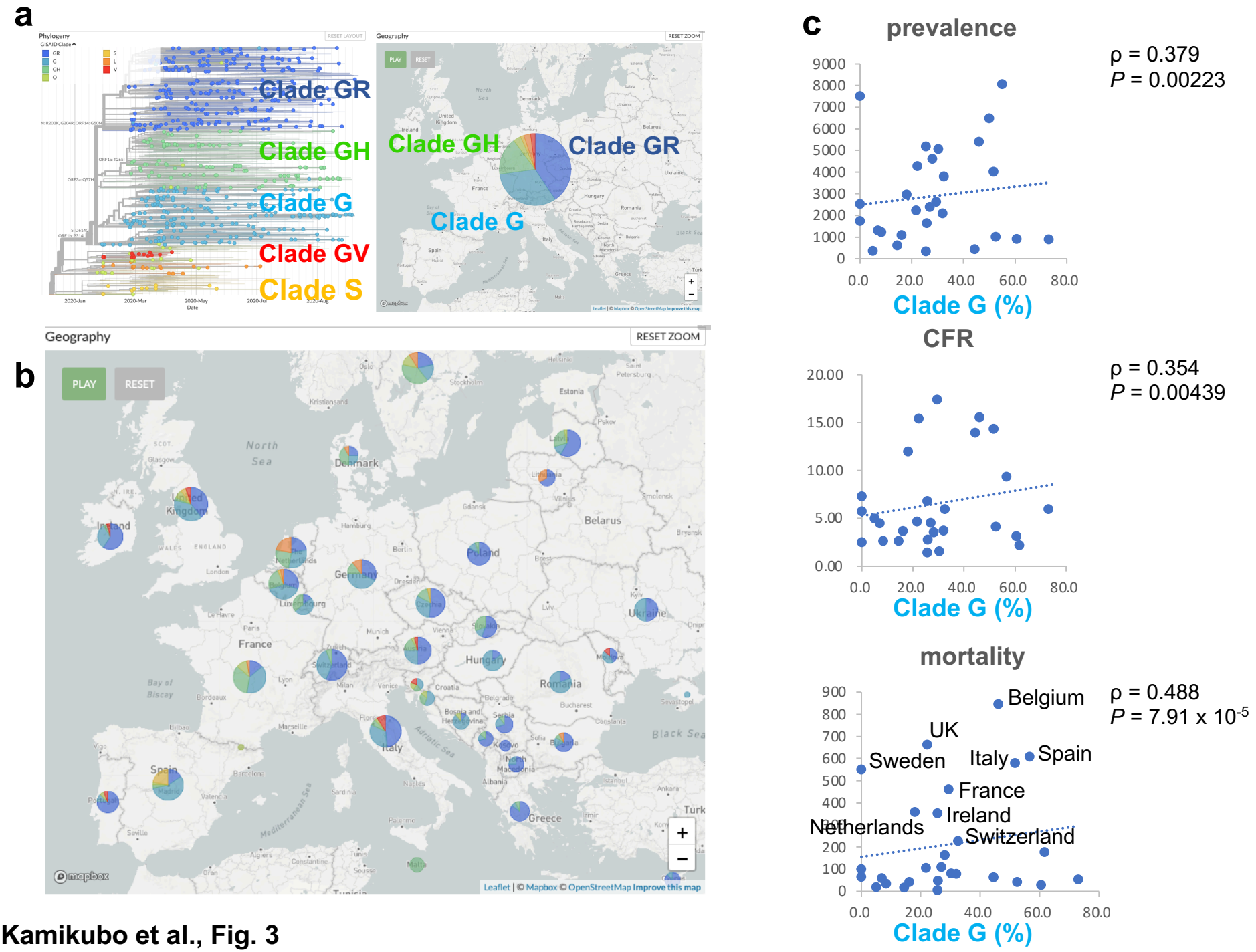
CFR are shown in red. **b, "**Epidemic curve" of SARS-CoV-2 mutations associated with viral spread. The number of bifurcated mutations for each month was plotted. Mutations at the same site of the protein that occurred in other parts of the phylogenetic tree were counted as different mutation events.



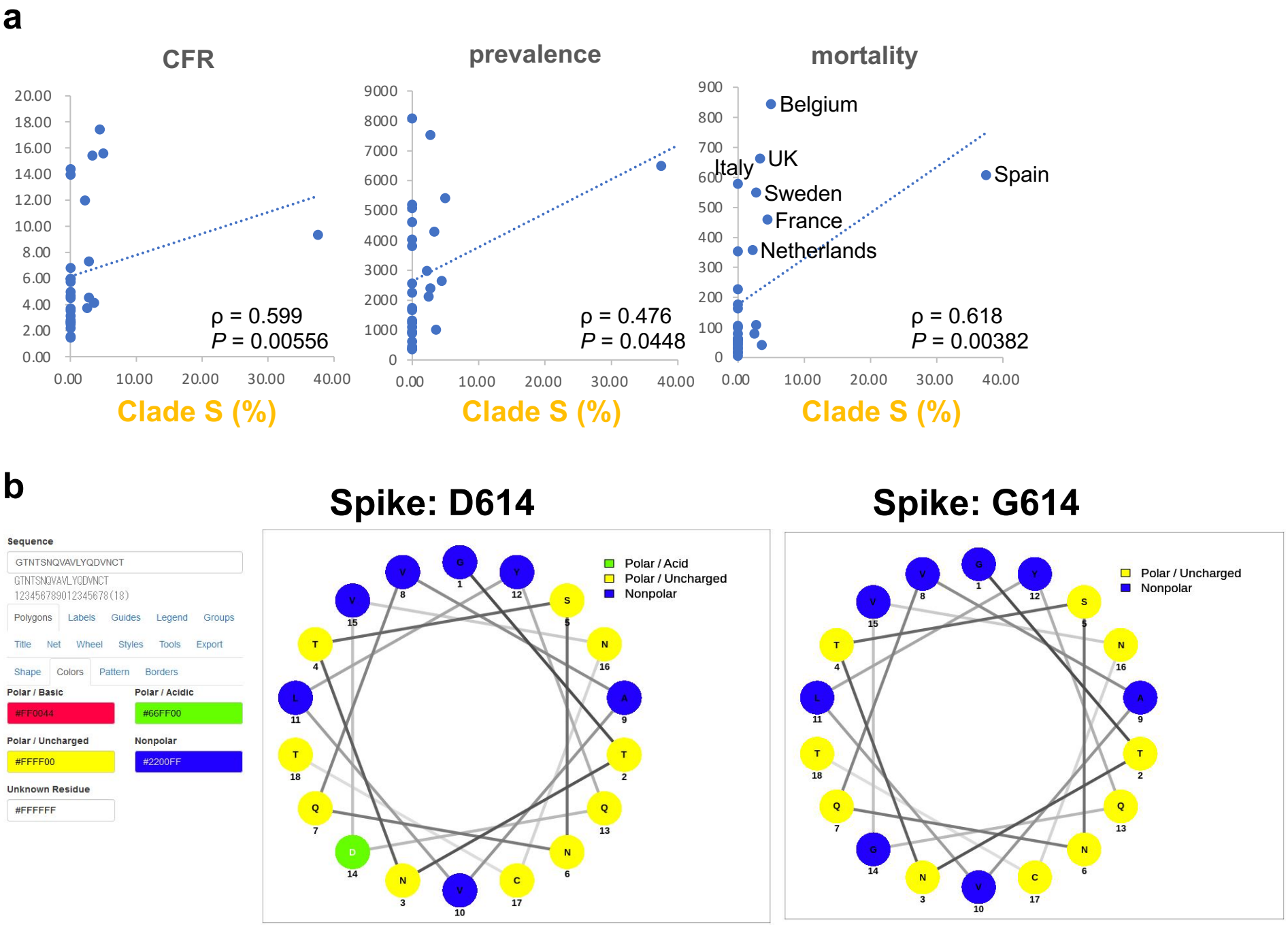
Kamikubo et al., Fig. 1



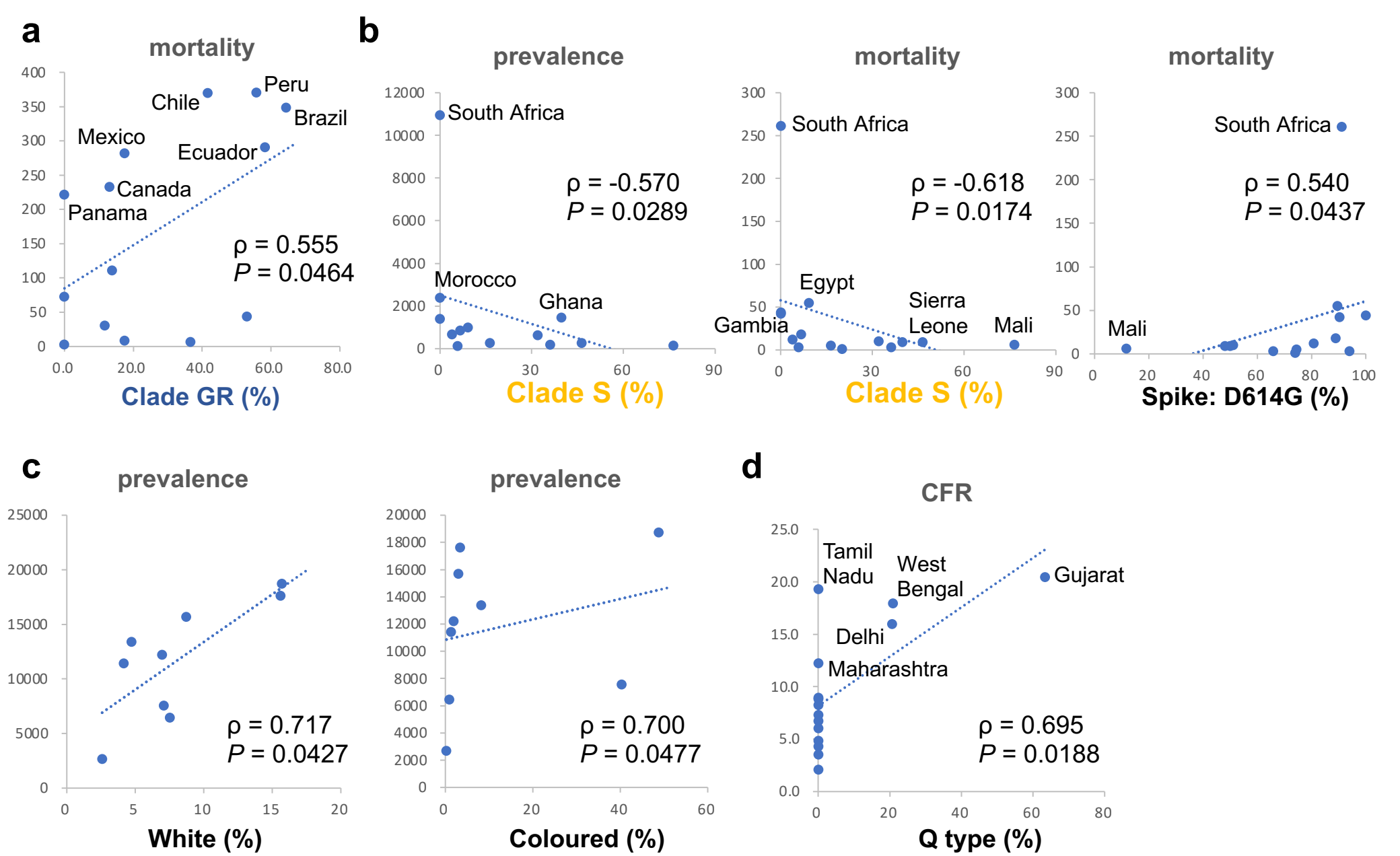
Kamikubo et al., Fig. 2



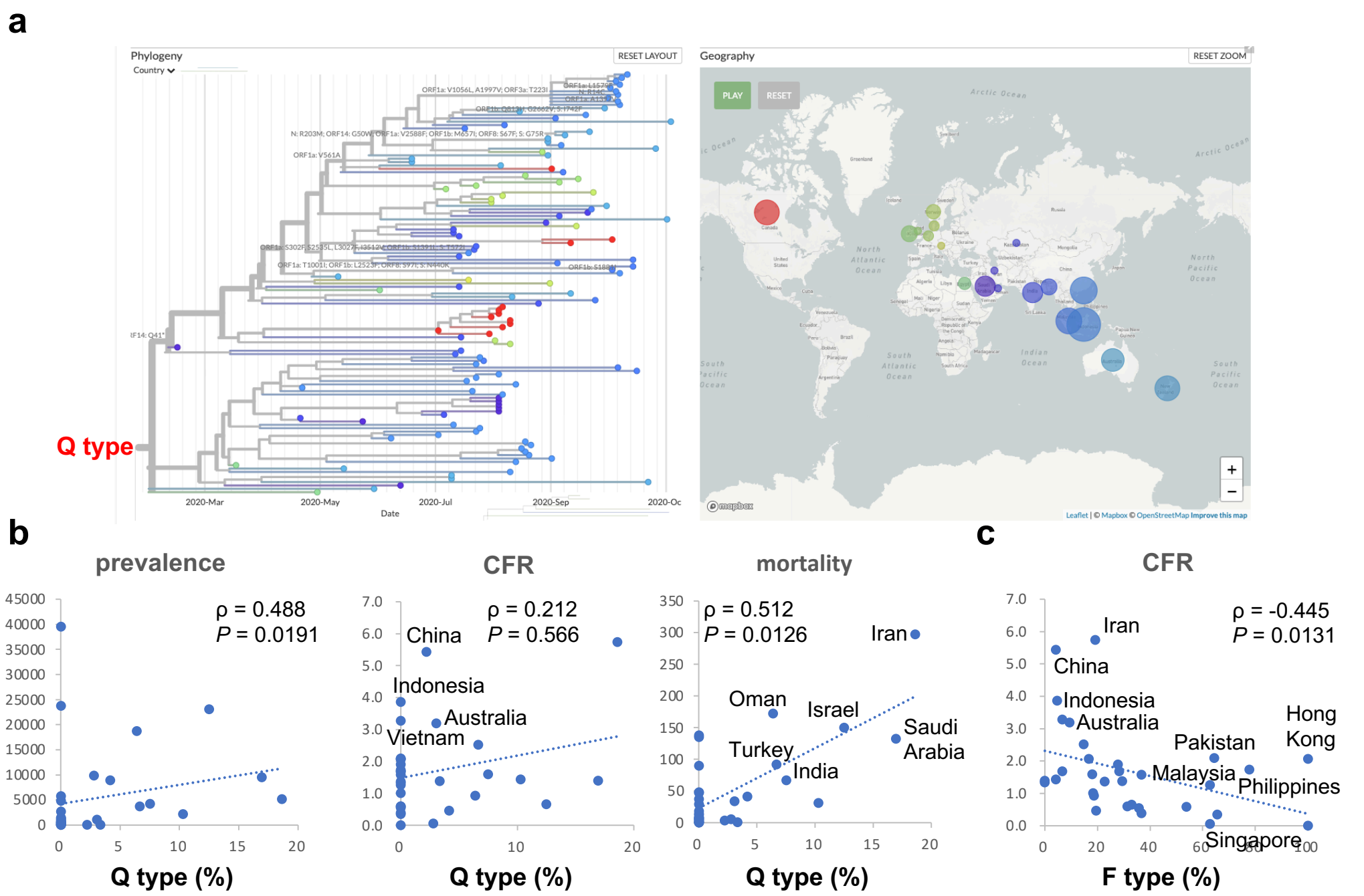
Kamikubo et al., Fig. 3



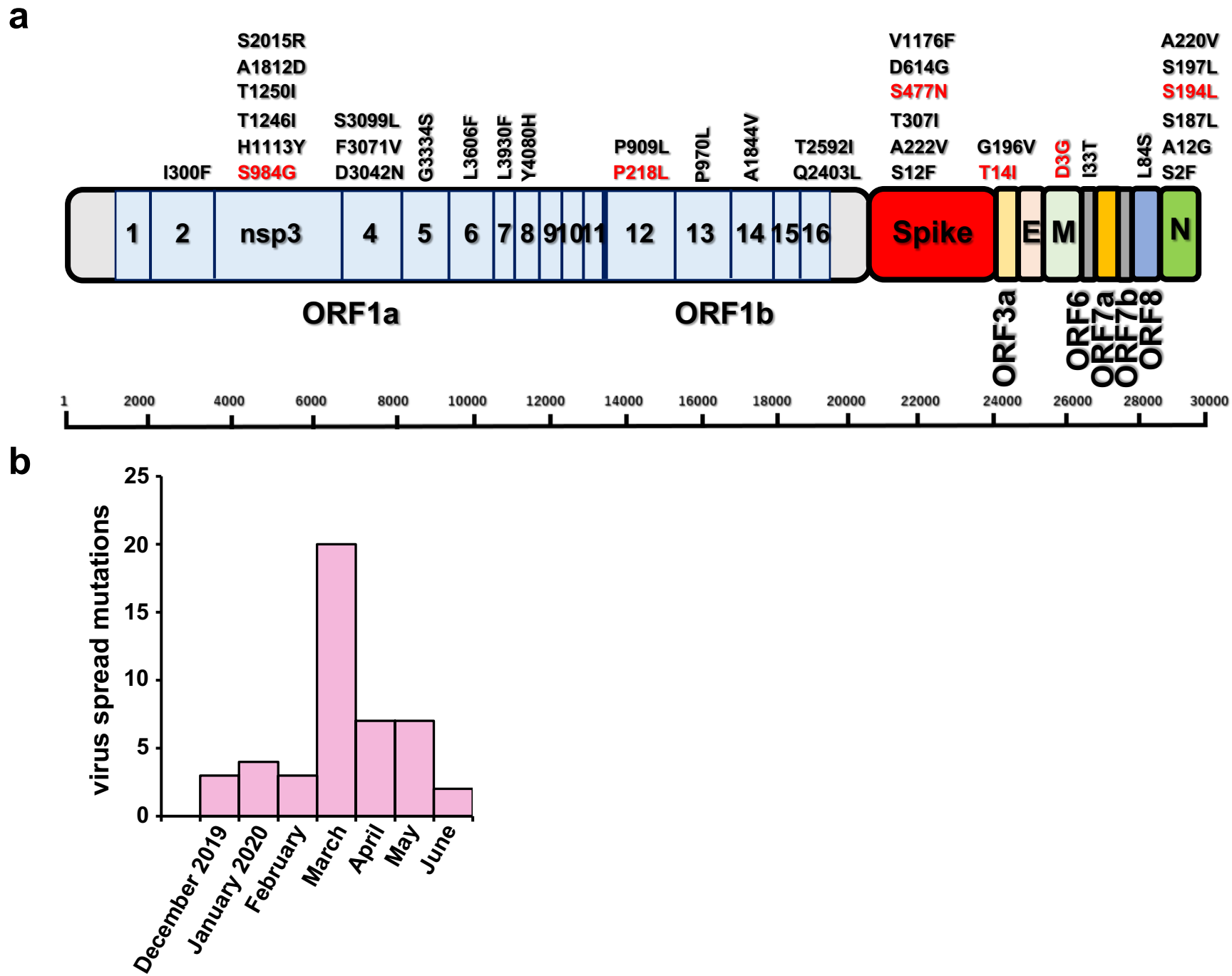
Kamikubo et al., Fig. 4



Kamikubo & Takahashi, Fig. 5



Kamikubo & Takahashi, Fig. 6



Kamikubo et al., Fig. 7