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

Yasuhiko Kamikubo, Toshio Hattori, Atsushi Takahashi

Department of Human Health Sciences, Graduate School of Medicine, Kyoto University, Sakyo-ku, Kyoto, Japan (Y Kamikubo MD); Graduate School of Health Science Studies and Research Institute of Health and Welfare, Kibi International University, Takahashi, Okayama, Japan (T Hattori MD, A Takahashi MD)

Correspondence to:

Dr Atsushi Takahashi, Graduate School of Health Science Studies, Kibi International University, Takahashi, Okayama 716-8508, Japan

tel. +81-866-22-9204

atakah7@kiui.ac.jp

Summary

Background The world has been plagued by complex epidemic waves of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that vary from region to region, leaving the end of the pandemic unpredictable. Epidemiology is expected to capture the current and future states of global epidemics, but available data were insufficient to constrain many possibilities.

Methods 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.

Findings A multifaceted analyses of the epidemic trends and their relationship to virus genotypes, regional herd immunity, population density, and race have provided the following lessons: (1) Spread of infection can occur in areas where viral fitness has increased due to mutations of viral proteins involved in intracellular proliferation; (2) Immunity against previously prevalent SARS-CoV-2 subtypes either protects against the spread of new subtypes or exacerbates coronavirus disease 2019 (COVID-19); (3) Immune evasion due to viral mutations leads to spread and severity of infection; (4) Viral competition with coexisting subtypes can prevent the outbreak of new invading subtypes; (5) Dense and crowded living environment can lead to viral spread and high lethality; (6) Racial and social disparities can lead to serious outbreaks of infection; (7) Viral mutations are epidemic, and there seems to be an upper limit to viral mutations that enable natural selection.

Interpretation These findings provide biological insights into how evolution through viral mutations shape the epidemic, as well as an overview of how host-side immune responses and social conditions have yielded the current pandemic outcomes. These insights and overviews help us predict the future and develop effective countermeasures.

Funding The Japan Society for the Promotion of Science.

Introduction

The pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1] is creating a complex wave of epidemics that varies from region to region. 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.

Epidemiology is expected to predict the future of infection, but there is too much uncertainty to narrow down from the many possibilities.[2,3] Previously, we developed epidemiological tools using the influenza epidemic curves.[4] Analyses of epidemic trends in Japan revealed that multiple viral subtypes had invaded. 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). 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.

We have also presented evidence for the involvement of Spike: D614G mutation in immune pathogenesis including ADE.[4] To our knowledge, it was the first paper focusing on Spike: D614G. Subsequent reports have revealed that this mutation also increases infectivity of the virus.[5,6] Although knowledge of SARS-CoV-2 gene mutations and a wealth of epidemiological information has been accumulated and published, there are few studies linking them.[7-9] Precise mathematical modelling has been performed to clarify the implications of mutations;[10] however, the usefulness of introducing ecological and geographical information[11] has not been fully tested. Here we carried out genomic epidemiological analyses[12] and conducted a multifaceted study that included ecological

and demographic factors, validating the proposed hypothesis[4] and revealing the global trends of SARS-CoV-2 epidemics.

Methods

Sources of Data

Data were obtained from websites including the Worldometer, Wikipedia, Governing, State & County Rankings, China Africa Research Initiative, [World economy newsletter], and the Ministry of Health and Family Welfare, Government of India. The GISAID database[13] 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.[11] Helix-wheel projections were created to analyse amphiphilic helix structures using the NetWheels projections maker.[14] Statistical analyses were performed with the use of the Statcel4 add-in package (OMS Publishing, Tokorozawa, Japan) for Microsoft Excel.

Role of the funding source

The funder of this study had no role in study design, data collection, analysis, interpretation, or writing of the report.

Results

To perform "genetic fingerprinting",[11] we referred to the Global Initiative on Sharing All Influenza Data (GISAID) website[13] for gene mutations that cause non-synonymous changes in viral proteins. The second wave of epidemic in Europe from the fall of 2020 has been mainly due to ORF10: V30L; N: A220V/ORF14: L67F; Spike: A222V variant, named GISAID Clade GV (fig. S1A).[7] To assess the epidemiological characteristics of Clade GV, we conducted an ecological study[11] comparing the frequency in the Spanish Communities (fig. S1B) with prevalence, CFR, and mortality. Prevalence was positively correlated with Clade GV frequency (fig. S1C), suggesting that the subtype is more transmissible than previously prevalent SARS-CoV-2.

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 major countries around the world (figure 1). CFR was high in Western Europe,[15] Mexico, Canada, and Ecuador, where the prevalence was not very high. In contrast, Chile, USA, Peru, and Brazil had high prevalence, but CFR was low. These indicate that there are regional factors that determine the spread and severity of COVID-19. In the epidemiological analysis of viral properties, it was considered necessary to stratify by region in order to eliminate these factors.

Europe

We first performed the ecological study of European countries (table S1). 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. Clade G was positively correlated with prevalence, CFR, and mortality (figure 2C). Clade GR had a weak positive correlation with mortality

(Spearman correlation coefficient $\rho = 0{\cdot}29$, $P = 0{\cdot}021$) and no significant correlation was observed with Clade GH ($\rho = 0{\cdot}028$, $P = 0{\cdot}83$). Clades G, GR, and GH share Spike: D614G mutation, which increases viral infectivity[5,6] and viral load.[9] Therefore, the results suggest that N: RG203KR/ORF14: G50N and ORF3a: Q57H mutations in Clade GR and GH reduced the virulence of Spike: D614G. However, analysis of these predominant Clades could not explain why CFR was high in Europe (figure 1).

We previously found that SARS-CoV-2 types S, K, and G invaded Japan in sequence and established partial herd immunity.[4] The study provided circumstantial evidence that the type S and G correspond to ORF8: L84S (Clade S) and Spike: D614G, respectively. Calculations using epidemiological parameters and Fermi estimates led to the following equation:

$F = 4{\cdot}7z - 166{\cdot}20y + 175{\cdot}58x$

where F is CFR and x, y, and z are exposure to type S, K, and G virus, respectively.[4] This formula predicts that previous exposure to type S virus leads to high CFR. Therefore, we plotted the relationship between Clade S and CFR in Europe. As predicted by the formula, CFR was positively correlated with Clade S (figure 3A). Multiple regression analyses (MRA) comparing the involvement of Clade G, GR, GH, and S revealed that only Clade S contributed strongly to mortality (standardized partial regression coefficient = $0{\cdot}452$, $P = 0{\cdot}0158$). Notably, except for Spain, Clade S accounts for less than 10% of SARS-CoV-2, indicating that virus-intrinsic causes cannot explain such high CFR.

Spike: D614 is located in a predicted antibody epitope,[16] forming a partially exposed structure resembling an amphiphilic helix.[4] The D614G mutation changes an acidic residue to a hydrophobic one, causing structural changes that alter antigenicity (figure 3B), which can convert the Clade S virus spike (D614) antibody to low affinity.

This conversion to low-affinity spike antibody[17] is expected to induce antibody-dependent enhancement (ADE).[18,19]

United States

Unlike Europe, no viral variants were found that significantly correlated with epidemiological parameters in US states. Instead, positive correlation with the proportions of non-Hispanic black (NHB) was noted in prevalence, CFR, and mortality (table S2 and fig. S2A).[20] Hispanic was positively correlated with prevalence and mortality.[21] Population densities were positively correlated with CFR and mortality (table S2 and fig. S2B). MRA suggested that NHB and Hispanic contributed more to the prevalence and mortality than population density (table S3), while population density seemed to be a major determinant of CFR. Interestingly, non-Hispanic American Indian was negatively correlated with CFR and mortality (table S2 and fig. S3). Overall, in USA, the living and social conditions indicated by population density and race[22] may have a major impact on the epidemic outcome (see Supplementary Text).

Latin America and Canada

Next, we analysed the effects of viral mutations in the countries of the Americas (table S4). Races was so influential in the US epidemic that USA was excluded for a stratified analysis. In contrast to Europe, only Clade GR was positively correlated with mortality (figure 4A). Therefore, we investigated whether new mutations that increase virulence were added to Clade GR. The high prevalence in Peru was associated with ORF1a: T1246I originating from Clade GR in Europe and spreading to Peru (fig. S4A). ORF6: I33T originated from Clade GR and spread to Brazil and Chile (fig. S5, A and C). Spike:

V1176F also originated from Clade GR in Brazil, with the additional spread-associated ORF1a: L3930F mutation (fig. S5B). N: S2F; ORF1a: T1250I; Spike: T307I originated from Clade GR in Chile with viral spread (fig. S5C). ORF1b: T2592I originated from Clade GR in Ecuador (fig. S4B). The emergence of new variants from Clade GR appeared to have contributed to the high prevalence (figure 1) and mortality (figure 4A) in Peru, Chile, Brazil, and Ecuador (see Supplementary Text).

CFR was higher in Mexico and Ecuador than other Latin American countries. To assess which variants boosted CFR in Mexico, ecological study was conducted in Mexican states. N: S194L originated from Clade G in Europe and spread to Mexico (fig. S6A). MRA showed that N: S194L was a major contributor to mortality (table S5). N: S194L tended to increase the prevalence, CFR, and mortality (fig. S6B). M: D3G originated from Clade G in Europe and spread to Ecuador (fig. S4, A and B). ORF1a: S984G originated from Clade G in Ecuador (fig. S4B). Since Clade G was associated with higher CFR in Europe (figure 2C), the N: S194L, M: D3G, and ORF1a: S984G from Clade G could account for high CFR in Mexico and Ecuador.

Canada had the second highest CFR after Mexico, so we searched for mutations that boosted CFR in the country. ORF3a: T14I was the main subtype only in Quebec (fig. S8B). Ecological studies comparing Canadian provinces were conducted on variants including Spike: D614G, ORF1a: T265I, ORF3a: T14I, Clade G, GH, and GR. MRA revealed a major contribution of ORF3a: T14I to prevalence, CFR, and mortality (table S6). In addition, more mutations have occurred downstream of ORF1a: T265I associated with viral spread (fig. S8B). ORF3a: T14I and these mutants from ORF1a: T265I appeared to explain the high mortality in Canada. Branching of numerous mutants from ORF1a: T265I was also observed in USA (fig. S7), possibly contributing to high

prevalence (figure 1). Therefore, in the Americas, excluding USA, whether a new virulent variant spread in that country seemed to determine the epidemic outcome.

Africa

Among African countries with GISAID data (table S7), Clade S was negatively correlated with prevalence and mortality (figure 4B), supporting the speculation that S type is an attenuated virus.[23] Spike: D614G correlated with increased mortality (figure 4B), consistent with increased viral infectivity.[5,6]

In the early epidemics in China, Clade S was the main type in areas other than Wuhan.[23] Due to many Chinese workers working in Africa, SARS-CoV-2 may have been introduced early. Therefore, we calculated the ratio of Chinese migrant workers in 2018 to each country's population. MRA showed that the Chinese worker ratio contributed more strongly to prevalence and mortality than Clade S (table S8). This result supports that the early introduction of attenuated coronaviruses by Chinese workers immunized Africans and prevented subsequent serious outbreaks (see Supplementary Text).[24]

South Africa had the highest prevalence among African countries. ORF1a: Y4080H originated from Clade GR in South Africa with viral spread (fig. S10A). ORF1a: T1246I from Clade GR, described above, also spread to South Africa. ORF1b: P970L from Clade G was also widespread in South Africa (fig. S10A). ORF1a: S3099L from Clade GR spread in The Gambia (fig. S10B). The emergence of ORF1a: Y4080H, ORF1a: T1246I, and ORF1a: S3099L from Clade GR may underlie the contribution of Clade GR to increased prevalence and mortality in the African region (table S8).

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 (figure 4C). Since whites tend to live in densely populated areas, we conducted MRA of race and population density. Whites were the main contributors to prevalence (table S9). In contrast, coloured races and population density contributed significantly to mortality (table S9).

India

As of November 2020, India had the second highest number of infections and the third highest number of deaths in the world. Q type (see below) spread to Gujarat, Delhi, and West Bengal. Q type was positively correlated with CFR in Indian states and Union Territory (figure 4D). MRA also showed that Q type contributed significantly to CFR (table S10). ORF1a: A1812D originated from Clade GR, spread to Maharashtra, Gujarat, and Telangana (fig. S11A), and appeared to be the major subtype in South India. MRA showed that ORF1a: A1812D contributed to prevalence and mortality (table S10) (see Supplementary Text). ORF3a: L46F, derived from Clade GR and spreading within Telangana, tended to lower CFR (fig. S11C) and contributed to lower mortality (table S10). F type (see below) (fig. S12A) contributed to prevalence (table S10) but did not contribute to mortality, probably because it tends to reduce CFR. It is also noteworthy that population density contributed to both prevalence and mortality (table S10). Therefore, in India, human congestion and immunity to spreading virus appeared to influence epidemic outcomes (see Supplementary Text).

Asia and Oceania

N: S194L/ORF14: Q41* originated from Clade GH and became widespread in Asia and Oceania (figure 5A). This variant, tentatively named Q type, was positively correlated

with prevalence and mortality (figure 5B). Consistent with ecological studies in Indian states (figure 4D), Q type tended to increase CFR (figure 5B). MRA showed that Q type was the major contributor to mortality (table S11). ORF1a: L3606F originated from Clade O (tentatively named F type) in the early stages of the epidemic and spread worldwide (fig. S12A). F type correlated with a decrease in CFR (figure 5C).

ORF1a: G3334S originated from Clade GR and spread to Oman (fig. S12B). This subtype accounts for 75% of Omani virus and may have contributed to high prevalence (figure 1). ORF1b: A1844V; ORF1a: S3884L originated from ORF1a: T265I in Clade GH and spread in Israel (figs. S9A and S13A). Another ORF1a: L3606F independently originated from ORF1a: T265I and spread to Israel (fig. S13B). ORF1a: L3606F is frequently scattered in the phylogenetic tree (fig. S13C), suggesting convergent evolution.[12] N: S194L downstream of ORF1a: L3606F was also associated with viral spread (fig. S13B). Like USA and Canada (figs. S7 and S8B), various mutations downstream of ORF1a: T265I may have contributed to the increased prevalence in Israel (table S12).

Indonesia had the highest mortality in the East and Southeast Asia due to high CFR. 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 and Central Java (fig. S14A). ORF1a: D1532G originated from Clade L and its spread was limited to Papua Province (fig. S14B). MRA suggested that ORF1b: P218L and ORF1a: D1532G increased and decreased CFR, respectively (table S13). Therefore, the high CFR seemed to be due to new mutations occurring in Indonesia.

Mortality in Bangladesh and Australia was comparable to Indonesia. ORF1a: I300F originated from Clade GR and spread from Bangladesh to Australia with viral

expansion (fig. S15A). ORF1a: I300F were dominant in Bangladesh (fig. S15, B and C) and appeared to increase prevalence in the country.

In Australia, Spike: S477N occurred downstream of ORF1a: I300F from Clade GR (tentatively named N type), causing an infectious outburst confined to the country (figs. S15A and S16A). An ecological study of Australian states was conducted and MRA revealed that N type contributed to increased prevalence and mortality (table S14). N type tended to increase CFR (fig. S16B). Spike: S477N also occurred independently from Clade GH in Europe together with N: M234I, A376T; ORF1a: M3087I; ORF1b: A176S, V767L, K1141R, E1184D (tentatively named DN type) (fig. S17A), suggesting convergent evolution. Clade GV, DN type, and ORF1a: H1113Y derived from Clade G in Belgium (fig. S17B) seemed to be the three major variants in the second wave of European epidemic as of November 2020.

Hong Kong had viral spread associated with Spike: S12F and N: A12G/ORF9b: H9D mutations in Clade GR (fig. S18A). ORF1a: S3884L originated from ORF1a: T265I in Clade GH and spread to South Korea (fig. S18B), where outbreaks were associated with additional ORF1b: Q2403L (fig. S18B). Viral spread in Singapore was associated with domestically prevalent ORF1a: S2015R from F type (fig. S19A) and ORF1a: D3042N from Clade GR (fig. S19B). Clade GV invaded Hong Kong and Singapore but did not spread. Partial herd immunity established in East and Southeast Asia may suppress the spread of Clade GV. Alternatively, prevalent virus subtypes may compete with Clade GV.[12] Populated areas such as Hong Kong and Singapore may increase selective pressure due to virus competition. In contrast to Clade GV, Q type began to spread in Singapore (figure 5A and fig. S19A) and Hong Kong (figs. S18A and S19C), demonstrating that Q type can surpass either regional immunity or virus competition. In summary, herd

immunity appears to be different in the eastern and western parts of the Asia-Oceania region. Against this background, the spread of different viral subtypes in each country may determine the epidemic outcome.

Time, Place, Mutation

To infer the molecular mechanism through which viral protein changes affect pathological processes, we created a "spot map" of the genomic locations of SARS-CoV-2 mutations (figure 6A). Mutations associated with viral expansion were concentrated in proteins involved in viral growth[25] (see Supplementary Text), potentially improving viral fitness[12] in the cell. To understand the trends of mutations involved in viral expansion, we plotted an "epidemic curve" of the mutations. Viral spread mutations peaked in March 2020 and have since peaked out (figure 6B), suggesting that some mechanisms that peak the mutations have worked and the mutation epidemic is ending. However, it is crucial to continue monitoring whether new mutations will turn to an increasing trend.

Discussion

Global comparison of SARS-CoV-2 prevalence and CFR showed that high mortality was due to high CFR in Europe and high prevalence in Latin America. In high-prevalence countries, new protein mutations associated with virus spread 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.[12] However, new variants in high-prevalence countries and its absence in low-prevalence countries contradicts the founder effect. Convergent evolution[12] has been suggested for many mutations, which may improve viral fitness. The emergence of mutants at the onset of the epidemic, rather than the gradual development during viral spread, supports the causal role of mutants, not the mere result of viral replication.

A global analysis of SARS-CoV-2 genomes reported that recurrent mutations did not increase the number of descendants, arguing against positive selection of mutants.[8] However, as the present study has revealed, transmissible variants expand in niches that reflect population heterogeneity and avoidance of defensive herd immunity and viral competition. A collective analysis of the world's subtypes[8] does not reveal such a niche spread. Such context-sensitive propagation of the subtypes could only be visualized by scrutinizing the local spread.

This study has supported the presumption that D614G virus in Europe caused ADE by low-affinity conversion of antibodies in areas where the Clade S virus had been present (see Supplementary Text).[4] Alternatively, the antibodies may cause conditions like vaccine-associated respiratory disease (VAERD)[17,26] and antibody-dependent cellular phagocytosis (ADCP).[27] Without any other rational explanations, the results support our hypothesis that immunity to type S virus underlies the exacerbation of COVID-19 in

Europe.[4]

Epidemiological effects also support that viral protein mutations alter the host's immune responses. High CFR-associated mutations were concentrated in proteins involved in the regulation of innate immunity. Mutations associated with higher CFR and prevalence are present in antibody epitopes,[28] HLA-DR T cell epitopes,[29] and predicted HLA-DR and HLA Class I T cell epitopes[16] (see Supplementary Text), suggesting that the mutations potentially reduce antibody avidity and modulate cell-mediated immunity. In evolutionary terms, these variants represent pleiotropic mutations that affect both antigenic traits and disease life-history traits.[10] Thus, the mutations can contribute to immune evasion and the exacerbation of COVID-19 mediated by innate immune system depression,[30] adaptive immune system dysregulation,[31] and incoordination between cell-mediated and humoral immunity.[32]

Taken together, the above considerations indicate that viral mutations drive the SARS-CoV-2 epidemic (see Supplementary Text). History has shown that if an infectious disease with a basic reproduction number (R_0)[11] ~2 and serial interval ~4 freely spreads in the population, it will reach herd immunity in about 3 to 4 weeks,[33] as was the case in Wuhan, China.[34] Highly transmissible subtypes with a higher R_0 raise the herd immunity threshold.[11] 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, existing herd immunity will fail, and the epidemic will resume. The protracted 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 RNA polymerase and RNA editing.[8] However, even if mutations continue to occur, they are not selected unless they cause amino acid changes that enhance virus transmission. As the mutation epidemic curve shows, the emergence of new mutations associated with viral spread has peaked out, strongly suggesting that there is little room for mutations to make more fit viruses. Indeed, in the 2003 SARS epidemic, protein mutation rates gradually slowed, after which SARS-CoV expansion ceased.[35] Similarly, when new protein mutations disappear, herd immunity to the current virus will be established and the epidemic is likely to end. We predict that the pandemic will end by March 2021 when winter ends in the Northern Hemisphere and mucosal defences and immune responses to winter respiratory viruses are restored.[36] This study also suggested that "mild cold" subtypes are beginning to drive out "severe pneumonia" subtypes (see Supplementary Text). Immunity to common cold coronaviruses lasts only about 10 months.[2,26] Even if the epidemic ends in spring 2021 in the Northern Hemisphere, the immune system will weaken in fall 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.[37]

Contributors

YK and AT conceived the study. 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.

Declaration of interests

We declare no competing interests.

Data sharing

After publication, study data will be made available on reasonable request to the corresponding author.

Acknowledgments

YK was funded by the Japan Society for the Promotion of Science (17H03597, 16K14632, and JP17H01690). We thank N. Takebayashi for data on Chinese in Africa and India; Y. Ota, T. Sakamoto, R. Taniguchi for collecting data from GISAID.

References

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

2. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 2020; **368**(6493): 860-8.

3. Saad-Roy CM, Wagner CE, Baker RE, et al. Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. *Science* 2020; **370**(6518): 811-8.

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

5. Hou YJ, Chiba S, Halfmann P, et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* 2020; **370**(6523): 1464-8.

6. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 2020: https://doi.org/10.1038/s41586-020-2895-3.

7. Hodcroft EB, Zuber M, Nadeau S, et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv* 2020: https://doi.org/2020.10.25.20219063.

8. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat Commun* 2020; **11**(1): 5986.

9. Volz E, Hill V, McCrone JT, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* 2021; **184**(1): 64-75 e11.

10. Day T, Gandon S, Lion S, Otto SP. On the evolutionary epidemiology of SARS-CoV-2. *Curr Biol* 2020; **30**(15): R849-R57.

11. Giesecke J. Modern Infectious Disease Epidemiology. 3rd ed: CRC Press; 2017.

12. Lauring AS, Hodcroft EB. Genetic Variants of SARS-CoV-2-What Do They Mean? *JAMA* 2021.

13. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018; **34**(23): 4121-3.

14. Mól AR, Fontes W, Castro MS. NetWheels: Peptides Helical Wheel and Net projections maker. 2020. http://lbqp.unb.br/NetWheels/.

15. O'Driscoll M, Ribeiro Dos Santos G, Wang L, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* 2020: https://doi.org/10.1038/s41586-020-2918-0.

16. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* 2020; **27**(4): 671-80 e2.

17. Bournazos S, Gupta A, Ravetch JV. The role of IgG Fc receptors in antibody-dependent enhancement. *Nat Rev Immunol* 2020; **20**(10): 633-43.

18. Eroshenko N, Gill T, Keaveney MK, Church GM, Trevejo JM, Rajaniemi H. Implications of antibody-dependent enhancement of infection for SARS-CoV-2 countermeasures. *Nat Biotechnol* 2020; **38**(7): 789-91.

19. Iwasaki A, Yang Y. The potential danger of suboptimal antibody responses in COVID-19. *Nat Rev Immunol* 2020; **20**(6): 339-41.

20. Price-Haywood EG, Burton J, Fort D, Seoane L. Hospitalization and Mortality

among Black Patients and White Patients with Covid-19. *N Engl J Med* 2020; **382**(26): 2534-43.

21. Vahidy FS, Nicolas JC, Meeks JR, 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* 2020; **10**(8): e039849.

22. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; **584**(7821): 430-6.

23. Lu J, Cui J, Qian Z, et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review* 2020; **7**(6): 1012-23.

24. Mbow M, Lell B, Jochems SP, et al. COVID-19 in Africa: Dampening the storm? *Science* 2020; **369**(6504): 624-6.

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

26. Sariol A, Perlman S. Lessons for COVID-19 Immunity from Other Coronavirus Infections. *Immunity* 2020; **53**(2): 248-63.

27. Pierce CA, Preston-Hurlburt P, Dai Y, et al. Immune responses to SARS-CoV-2 infection in hospitalized pediatric and adult patients. *Sci Transl Med* 2020; **12**(564): https://doi.org/10.1126/scitranslmed.abd5487.

28. Shrock E, Fujimura E, Kula T, et al. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science* 2020; **370**(6520): https://doi.org/10.1126/science.abd4250.

29. Nelde A, Bilich T, Heitmann JS, et al. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat Immunol* 2021; **22**(1): 74-

85.

30. Hadjadj J, Yatim N, Barnabei L, et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* 2020; **369**(6504): 718-24.

31. Lucas C, Wong P, Klein J, et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* 2020; **584**(7821): 463-9.

32. Rydyznski Moderbacher C, Ramirez SI, Dan JM, et al. Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. *Cell* 2020; **183**(4): 996-1012 e19.

33. Crosby AW. America's Forgotten Pandemic: The Influenza of 1918: Cambridge University Press; 2003.

34. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* 2020; **382**(13): 1199-207.

35. Chinese SMEC. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* 2004; **303**(5664): 1666-9.

36. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. *Annu Rev Virol* 2020; **7**(1): 83-101.

37. Lavine JS, Bjornstad ON, Antia R. Immunological characteristics govern the transition of COVID-19 to endemicity. *Science* 2021.

Figure legends

***Figure 1*: Prevalence and CFR of COVID-19 in major countries of the world**

A scatter plot showing the relationship between prevalence (number of cases per 1,000,000 people) and CFR (%) in 67 countries as of 15 July 2020. Countries with few PCR tests (less than 1300 tests per 1,000,000 people) or uncertain statistics were excluded because countries with inadequate population surveys would confuse the analysis. The countries in Europe, Americas, Africa, the Middle East, and Asia/Oceania are shown in green, magenta, light blue, purple, and brown, respectively.

***Figure 2*: Epidemic outcome in Europe dominated by SARS-CoV-2 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 (number of deaths per 1,000,000 people) in Europe as of 15 July 2020. ρ: Spearman correlation coefficient.

***Figure 3*: 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.

***Figure 4*: 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 (N: S194L/ORF14: Q41* from Clade GH) was positively correlated with CFR in Indian states and Union Territory as of 23 September 2020. ρ: Spearman correlation coefficient.

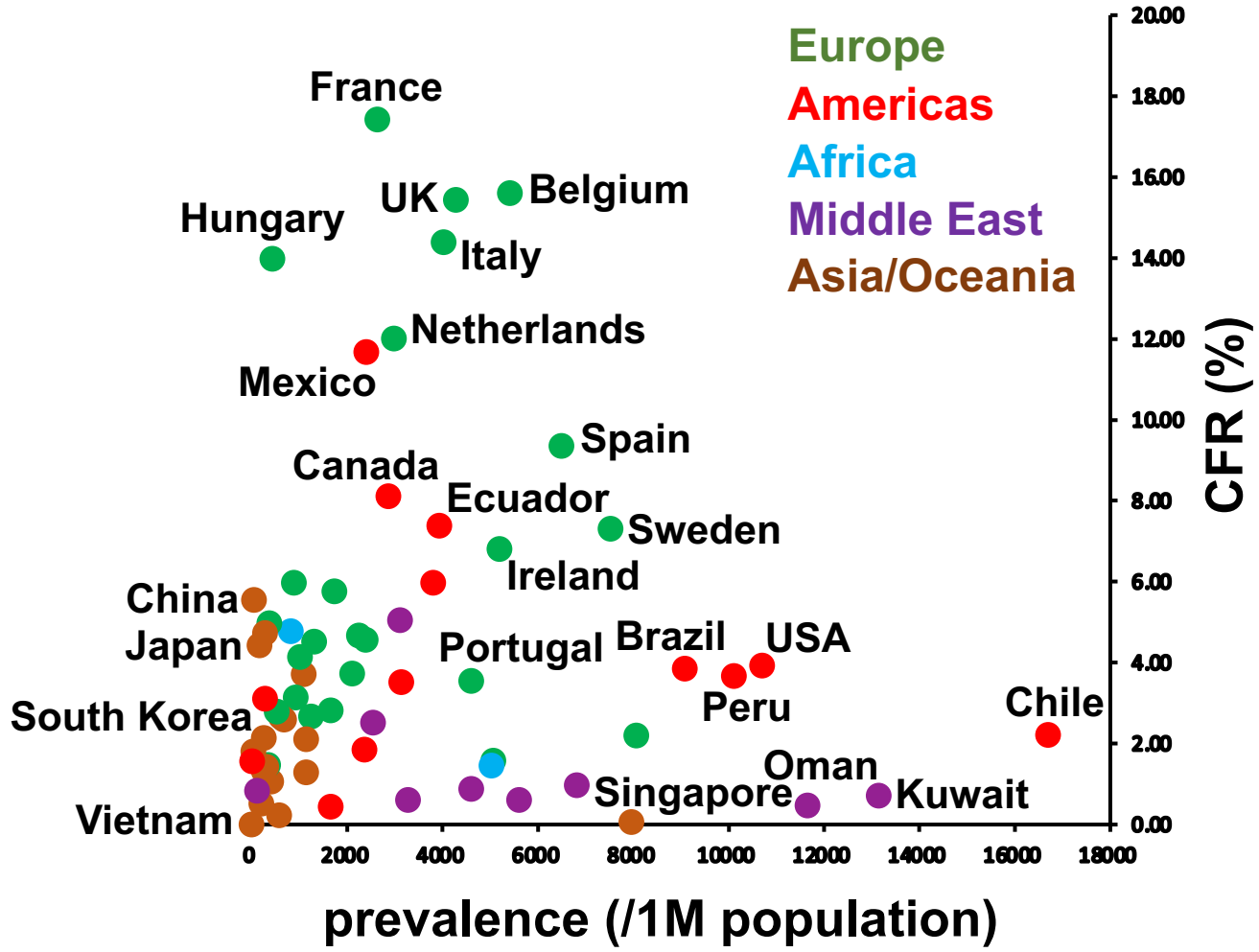
***Figure 5*: SARS-CoV-2 Q type in Asia and Oceania associated with increased mortality**

(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 colour 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) SARS-CoV-2 F type was negatively correlated with mortality in Asian countries. ρ: Spearman correlation coefficient.

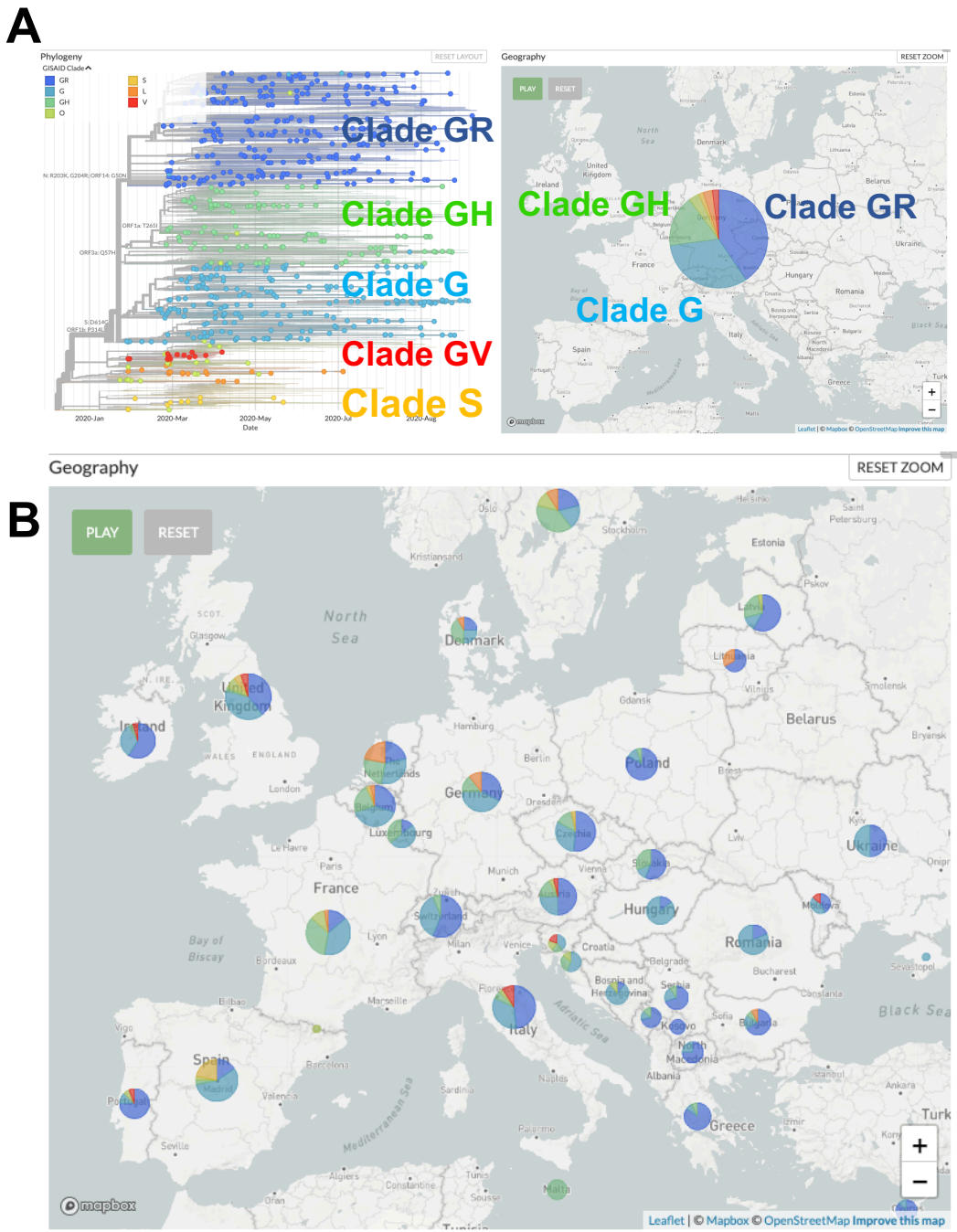
***Figure 6*: Epidemiology of SARS-CoV-2 mutations**

(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

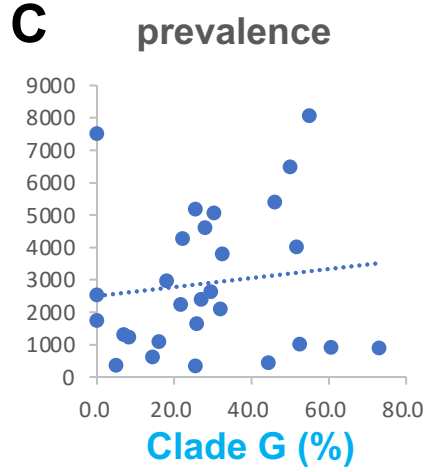
mutation events.



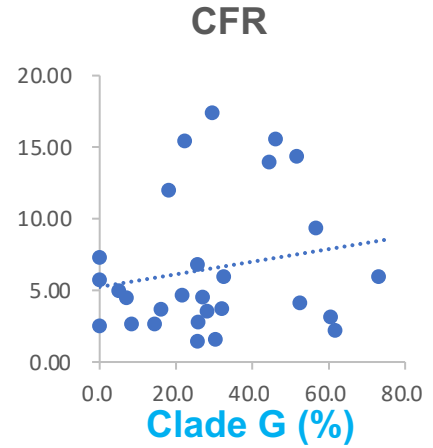
Kamikubo et al., Figure 1



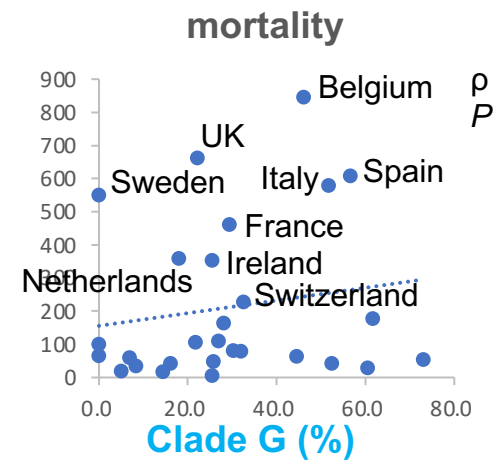
C



ρ = 0.379
P = 0.00223

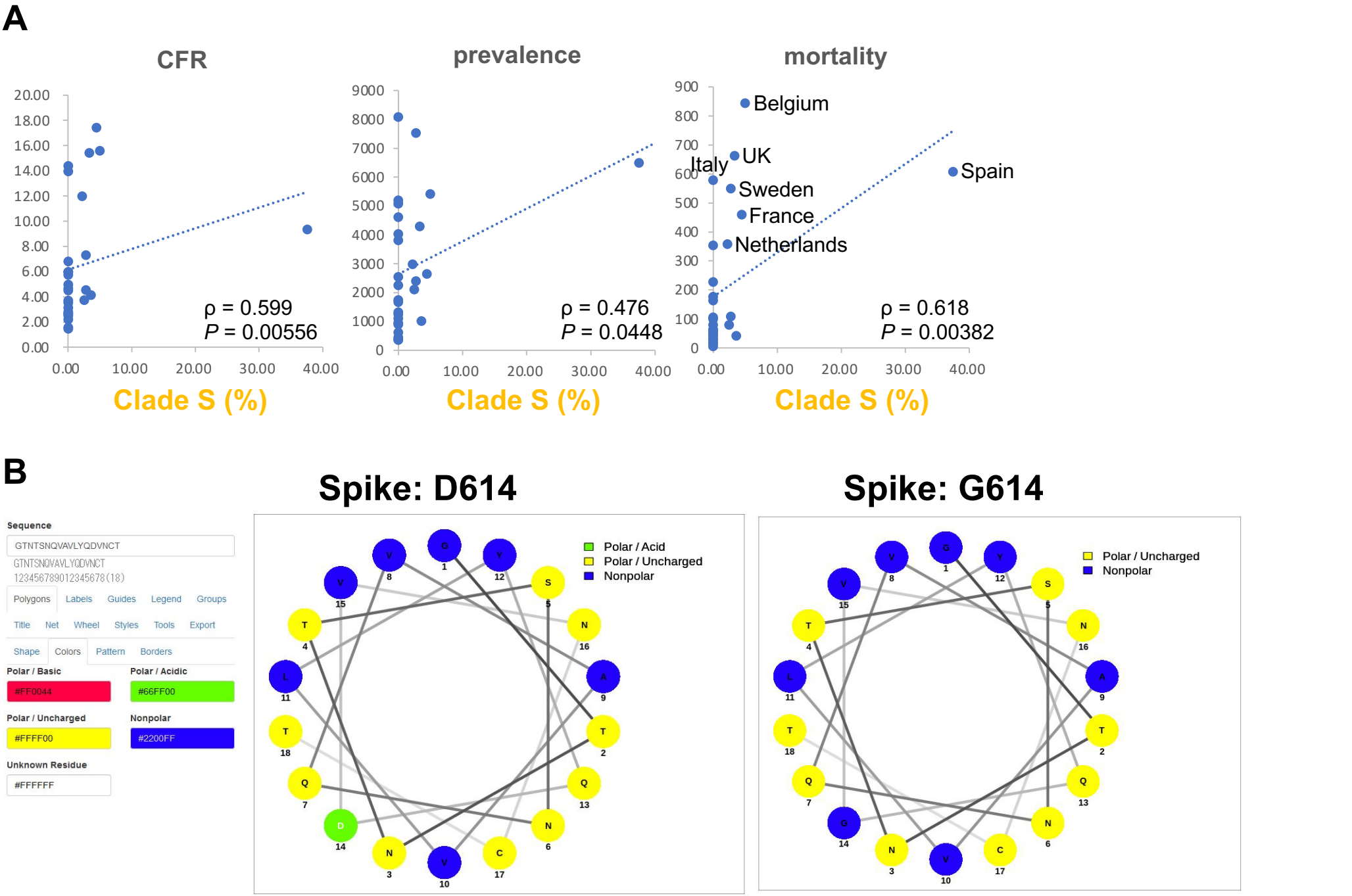


ρ = 0.354
P = 0.00439

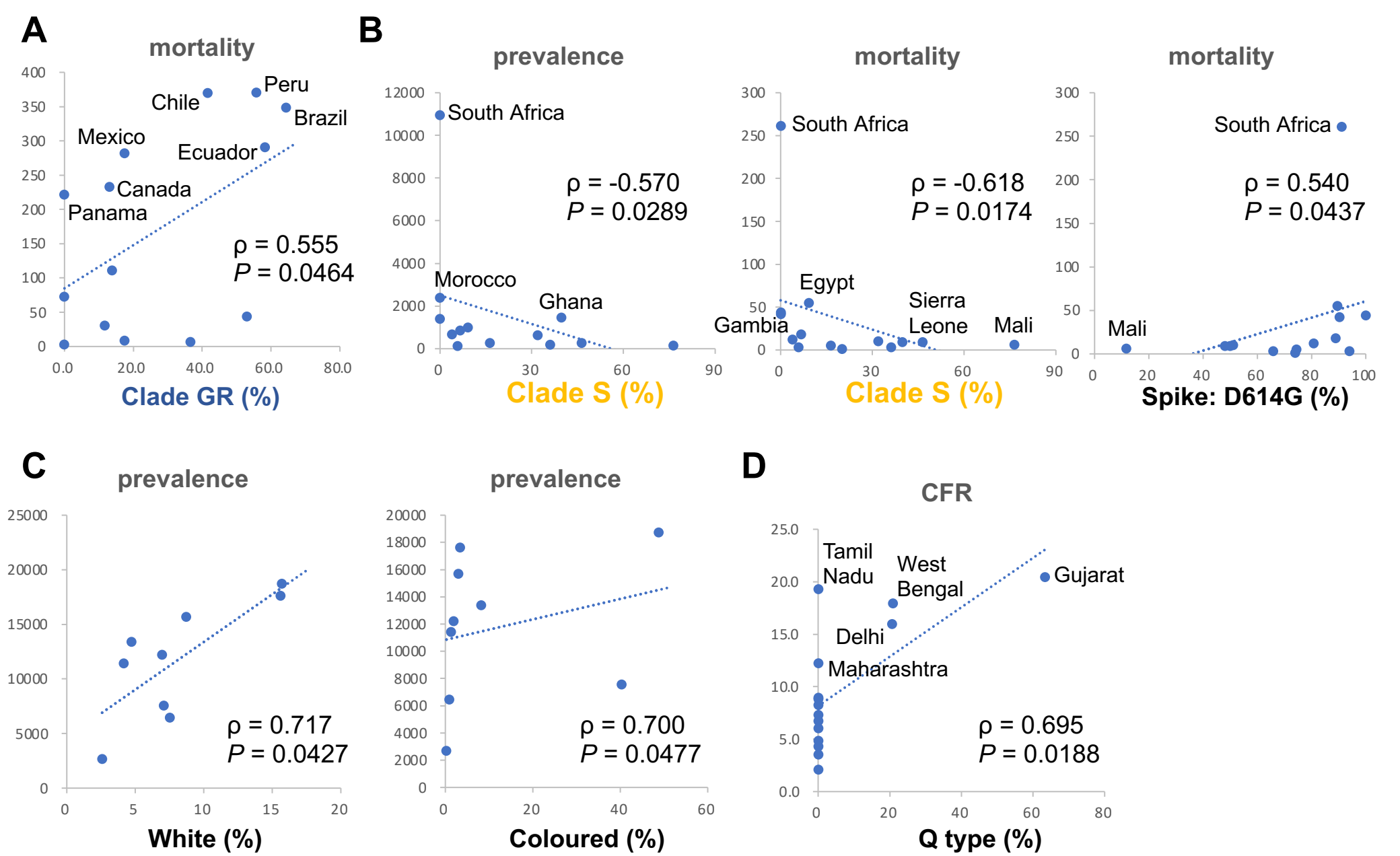


ρ = 0.488
P = 7.91 x 10^{-5}

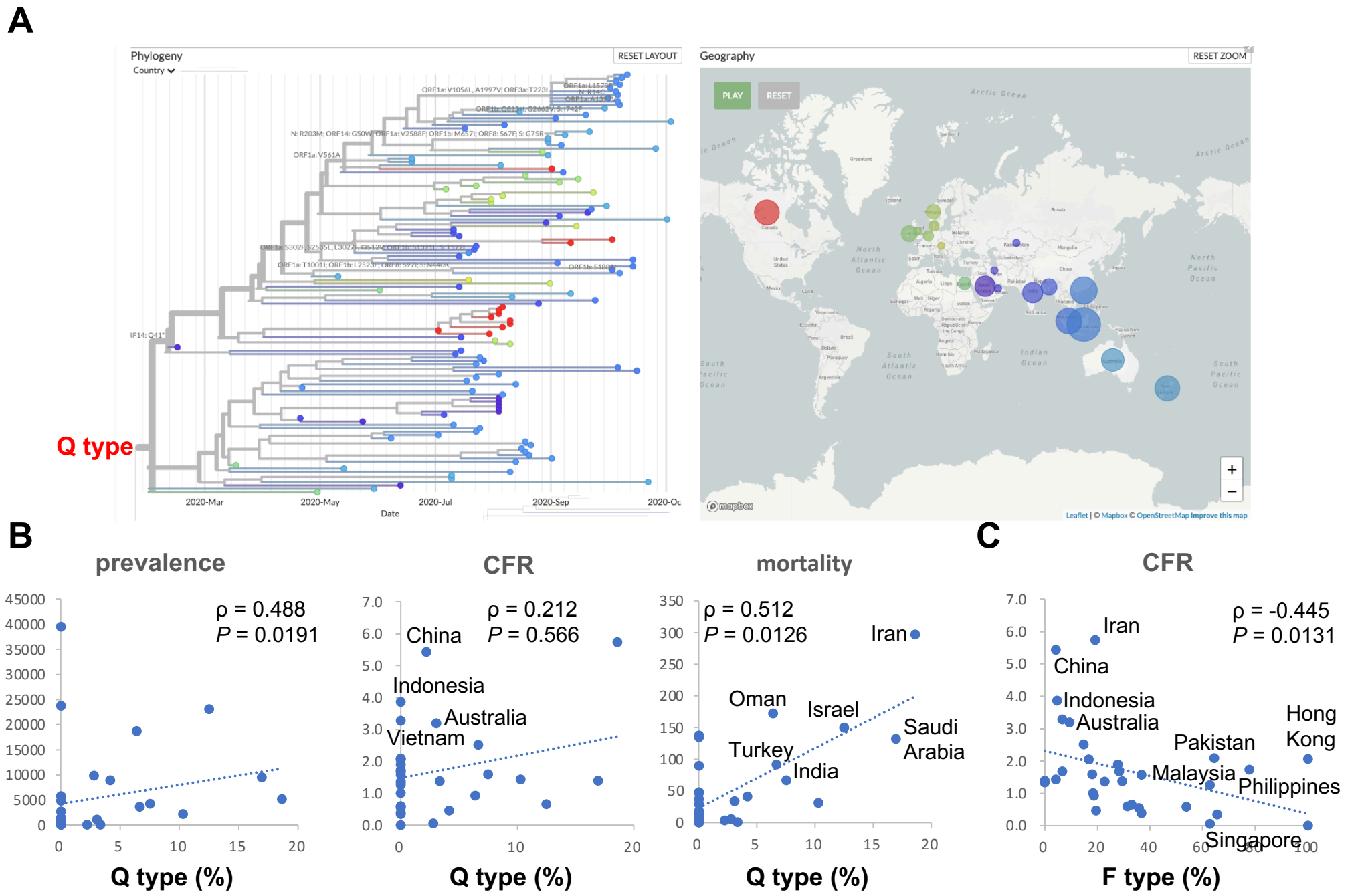
Kamikubo et al., Figure 2



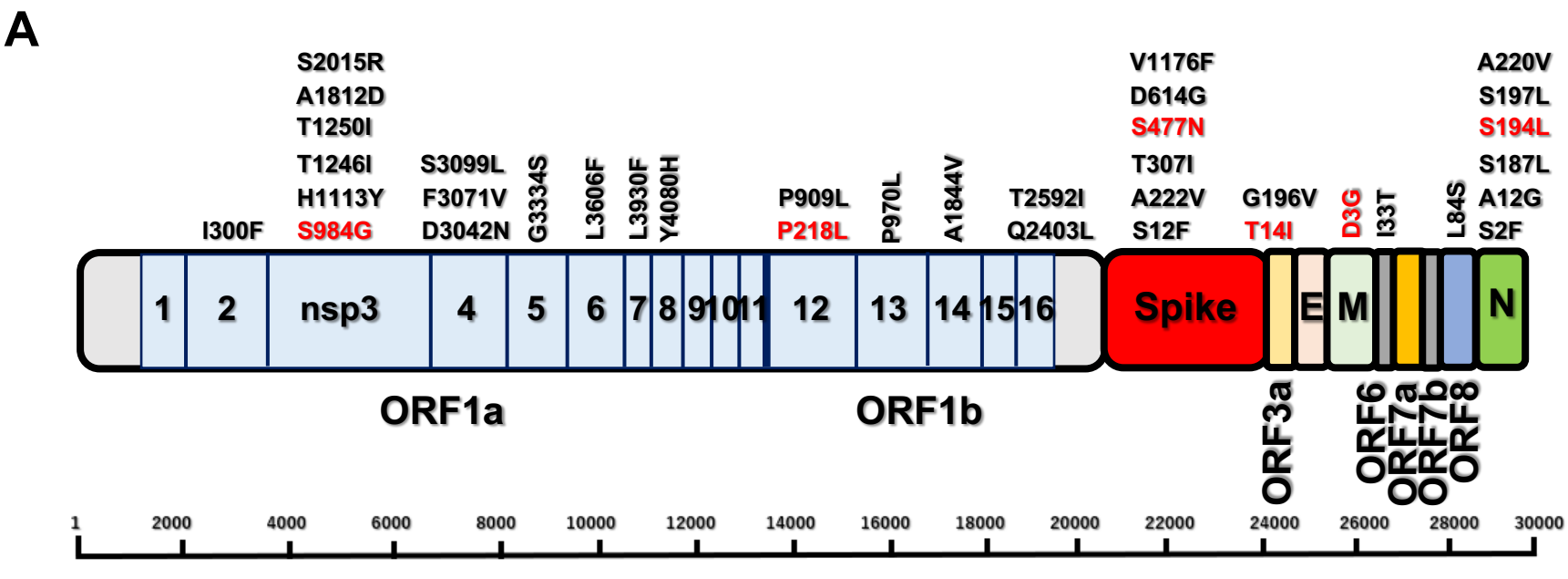
Kamikubo et al., Figure 3

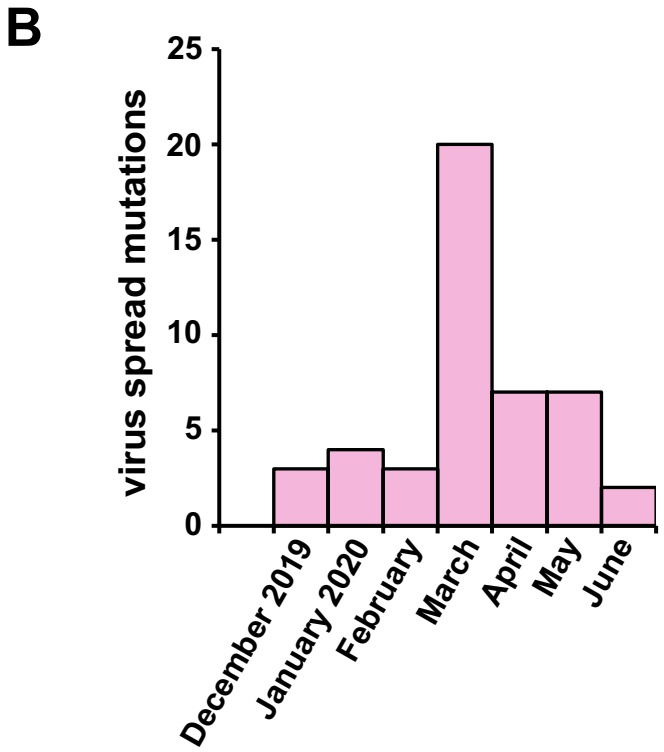


Kamikubo et al., Figure 4



Kamikubo et al., Figure 5





Kamikubo et al., Figure 6