Three adjacent nucleotide changes spanning two residues in SARS-CoV-2 nucleoprotein: possible homologous recombination from the transcription-regulating sequence

Running title: New motif within SARS-CoV-2 nucleocapsid

Shay Leary[1*], Silvana Gaudieri[1,2,3*], Abha Chopra[1], Suman Pakala[3], Eric Alves[2], Mina John[1,4], Suman Das[3], Simon Mallal[1,3^], Elizabeth Phillips[1,3^]

[1]Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, Western Australia, Australia

[2]School of Human Sciences, University of Western Australia, Crawley, Western Australia, Australia

[3]Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[4]Department of Clinical Immunology, Royal Perth Hospital, Perth, Western Australia, Australia

[*] Equal contribution

[^] Equal contribution

Corresponding Authors

Dr. Silvana Gaudieri

School of Human Sciences, University of Western Australia

M309, 35 Stirling Hwy, Crawley, Western Australia 6009

Australia

Tel: 61-8-6488 1096

Email: silvana.gaudieri@uwa.edu.au

Dr. Simon Mallal

Department of Medicine

Vanderbilt University Medical Center

Nashville, Tennessee, USA 37232-2605

Email: S.Mallal@vumc.org

Conflict of interest: None declared

Abstract

The COVID-19 pandemic is caused by the single-stranded RNA virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a virus of zoonotic origin that was first detected in Wuhan, China in December 2019. There is evidence that homologous recombination contributed to this cross-species transmission. Since that time the virus has demonstrated a high propensity for human-to-human transmission. Here we report two newly identified adjacent amino acid polymorphisms in the nucleocapsid at positions 203 and 204 (R203K/G204R) due to three adjacent nucleotide changes across the two codons (i.e. A**GG G**GA to A**AA C**GA). This new strain within the LGG clade may have arisen by a form of homologous recombination from the core sequence (CS-B) of the transcription-regulating sequences of SAS-CoV-2 itself and has rapidly increased to approximately one third of reported sequences from Europe during the month of March 2020. We note that these polymorphisms are predicted to reduce the binding of an overlying putative HLA-C*07-restricted epitope and that HLA-C*07 is prevalent in Caucasians being carried by >40% of the population. The findings suggest that homologous recombination may have occurred since its introduction into humans and be a mechanism for increased viral fitness and adaptation of SARS-CoV-2 to human populations.

Keywords: COVID-19; SARS-CoV-2; homologous recombination; viral polymorphism; adaptation, HLA

Background

Evidence of viral adaptation to selective pressures as it spreads among diverse human populations has implications for the ongoing potential for changes in viral fitness over time, which in turn may impact transmissibility, disease pathogenesis and immunogenicity. Geographic differences in viral sequence diversity and epidemiological profiles of disease are likely to reflect the spread of founder viruses, which first entered different SARS-CoV-2 naïve populations. However, the extent to which selection pressures operating within those populations also impact SARS-CoV-2 diversity is currently not known. Functional effects of new genetic changes need to be considered in ongoing public health measures to contain infection around the world and in the development of universal vaccines and antiviral therapy. Here we describe a new emerging strain of SARS-CoV-2 within the LGG clade that appears to be the result of a homologous recombination event that introduced three adjacent nucleotide changes spanning two residues of the nucleocapsid protein. That strain expanded rapidly in Europe in March 2020. This protein forms an integral part of the virus life-cycle and is known to be highly immunogenic.

Main text

Newly emerging strain from Europe with linked variations in the nucleocapsid

We utilized publicly available SARS-CoV-2 sequences from the GISAID database (www.gisaid.org; see Supplementary Material) to identify polymorphisms arising in global circulating forms of the virus in relation to region and time of collection. Of the 2336 circulating strains examined in this study there was limited variation across the genome with only 12 amino acid polymorphisms present in >5% of the deposited sequences (Supplementary Table 1). Of these polymorphisms, three were the

polymorphisms L84S in ORF8, D614G in surface glycoprotein (S) and G251V in NS3 (ORF3a) that mark the major worldwide clades S, G and V, respectively. Two newly identified adjacent polymorphisms (R203K and G204R) in the nucleocapsid protein occur in approximately 13.4% of deposited strains and form one of the main strains emerging from Europe (Figure 1A). Other common polymorphisms include Q57H in NS3, T85I in NSP2, L37F in NSP6, P323L in the RNA-dependent RNA polymerase, T175M in the membrane glycoprotein and P504L and Y541C in the helicase. Current low frequency polymorphisms at <5% of deposited SARS-CoV-2 sequences include S193I in the nucleocapsid, H93Y in NS3, and the following polymorphisms V378I, G392D, I739V, P765S, A876T, F3071Y, G3278S and K3353R in ORF1ab (Supplementary Table 1).

The polymorphisms are present in strains sequenced using different next generation sequencing (NGS) platforms (e.g. nanopore, Illumina) and the Sanger-based sequencing method making it unlikely that the new changes are sequence or alignment errors. In addition, different laboratories around the world have deposited sequences with these polymorphisms in the database and examination of individual sequences in the region does not find obvious insertions/deletions likely representing alignment issues or homopolymer slippage.

Adjacent amino acid polymorphisms due to three adjacent nucleotide changes in the nucleocapsid

For the two newly identified adjacent polymorphisms in the nucleocapsid at positions 203 and 204, there were no strains in the database that had only one of the two changes. The SARS-CoV-2 sequences deposited into the GISAID database are

consensus strains predominantly generated from NGS platforms that can typically identify low frequency variants. We did not have access to the original sequence files from the contributing laboratories in order to assess if there was evidence of strains that harbored only one of the polymorphisms at lower frequencies. However, no circulating strain has so far been captured that contains only one of the two nucleocapsid polymorphisms as the consensus sequence.

The rapid emergence of these closely linked polymorphisms in viruses may reflect strong selection pressure on this region of the genome in which the original mutation incurred a replicative capacity, or other fitness cost, which could be restored by a linked compensatory mutation. Evidence for such adaptations with closely linked compensatory mutations are known to occur under host immune pressure as is well established for other adaptable RNA viruses such as HIV [1,2] and Hepatitis C virus (HCV) [3]. These viruses have such a high rate of viral replication and error-prone reverse transciptase that a massive swarm of viral variants with ongoing recombination between residues is generated continuously. As a result selection pressure exerted by immune responses or other selective pressures effectively operate on each separate residue independently [4]. In contrast, coronaviruses encode proof-reading machinery and have a propensity to adapt by homologous recombination between viruses rather than classic step-wise individual mutations driven by selective pressures operating on single viral residues. This, together with the routine nature of their cross-species transmission [5], led Graham and Baric [6] to presciently warn in 2010 that it was a matter of when, rather than if, a pathogenic coronavirus pandemic would occur in humans. Also of note, the phenomena of compensatory fixation has been described in the area of HIV antiviral resistance in which the linked mutations cannot

revert to wild type when the selective pressure is removed as the virus cannot negotiate the fitness valley to return to its previous optimal state [7]. We therefore predict that the K203/R204 (AAA CGA) change is likely to remain fixed and intermediates to the wild type are unlikely to be found. It will be critical to determine if the introduction of the AAACGA motif results in a replicative or other fitness cost to the virus, creates an alternative subgenomic mRNA transcript or RNA secondary structure or increases nucleocapsid activity as this could indicate that there may be viral attenuation as passage occurs globally through populations of diverse immunogenetic background.

As further evidence of the likelihood of a homologous recombination event, the R203K polymorphism involves a two-step process from AGG to AAA. However, strikingly, the position shows no evidence to date of alternative codon usage, all viral strains that contain an R at this position have the AGG codon, and similarly those strains with a K (lysine) at this position have the AAA. Circulating strains that contain the intermediate codon as the consensus resulting from a single step to the AGG lysine amino acid (i.e. the arginine codon AGA or the lysine codon AAG) appear completely absent among captured strains to date.

To identify possible viral sources for homologous recombination with SARS-CoV-2, we initially performed a search of the motif in the nucleocapsid in related beta coronaviruses from human and other species in the public databases and only found the presence of the R203/G204 combination. We performed a similar search in our metatranscriptome data generated from a cohort study consisting of 65 subjects of which 43 had acute respiratory infections and 22 were asymptomatic. From the data

we assembled near complete and coding complete viral genomes of the Coronavirus (NL63 - alpha, OC43 - beta, 229E - alpha), RSV (A, B), Rhinovirus (A, B, C), Influenza (A - H3N2), and Bocavirus family. None of the alpha coronaviruses had the R203/G204 or K203/R204 combination or indeed any variation at these sites (n=14; sequence depth >3000). We then performed a search for stretches of similarity using varying window sizes (>14bp including the motif; [8]) in all sequences. No significant hits were identified. However, the AAACGA motif overlaps with the core sequence (CS-B) of the transcription-regulating sequences [9] of SAR-CoV-2 itself and this motif has been found several times within SARS-CoV-2 at the end of the protein for ORF1ab, surface glycoprotein, envelope, ORF6, ORF7b and ORF8 (Supplementary Figure 1).

New nucleocapsid polymorphisms likely emerged from existing strain containing the clade defining L84 (ORF8), G614 (S) and G251 (NS3) amino acids

As many of the sequences have been generated using NGS technologies that likely reflect single strains across the viral genome, we examined whether the new polymorphisms occurred on specific existing viral genetic backgrounds. The K203/R204 polymorphisms so far only exist on strains that carry the following clade defining amino acids - L84 in ORF8, G614 in S and G251 in NS3 to form the combination (haplotype) KR-LGG (Figure 1A). A chronological analysis of the emergence of the different strains as defined by the five positions (and based on the collection date assigned to the deposited sequences) shows the transition between major forms (Figure 1B). The first sequences identified in China contained the RG-LDG combination that closely clusters with the SARS-CoV-2 strain RG-SDG that was first deposited in early January of Chinese origin and is present in both the bat

and pangolin coronaviruses and the previous SARS-CoV strain associated with the 2003 outbreak. These two strains of SARS-CoV-2 are now prevalent on most continents (Figure 1A) and the RG-LDG strain was also the dominant strain that was identified in passengers and crew abroad the Diamond Princess cruise ship docked in Japan immediately prior to the more global dissemination of cases worldwide. The RG-LDG, RG-SDG and RG-LDV strains were circulating in Asia before mid-January, with entry into the West coast of the US with the first identified case in Washington State January 21, 2020 and by March the RG-SDG strain had become widespread in North America.

The appearance of the RG-LGG and KR-LGG strains occurred in the database in late February in Europe and for the RG-LGG strain in the US by early-mid March. To date the viruses with the KR-LGG haplotype form a significant portion of circulating strains in the Netherlands and Iceland with smaller numbers in neighboring European countries and elsewhere. In the Netherlands the first confirmed case of Sars-CoV-2 infection was on February 28, 2020 in an individual who had recently visited the Lombardy region in Italy. Six cases that followed were also reported to be of linked either to travel to Lombardy or through contact with the first patient [10]. Similarly, in Iceland the first reported case was also on February 28, 2020 from a returning traveler from Northern Italy with five cases that followed linked to Verona, Italy [10]. It is therefore likely that strains harboring these polymorphisms exist in Northern Italy but there is currently limited sequence data deposited on the specific viral strains in the most affected regions. In further support of this expectation, sequences deposited from the Brazilian government indicate the likely source of virus for the small number of strains with the KR-LGG haplotype in Brazil is Northern Italy. Similarly, the

source for the KR-LGG strains in Chile, Mexico and Nigeria has also been indicated as Northern Italy.

As of March 31, 2020 there appears to be only a small proportion of strains with KR-LGG in the US, likely reflecting that deposited sequences have been mainly from the West coast of the US that experienced initial importation of Asian strains of SARS-CoV-2. It will be of great interest to see sequences from the East coast of the US given the early importation of SARS-CoV-2 from Northern Europe as well as Asia and the widespread community transmission that has followed (Figure 1A).

Interestingly, the M175 polymorphism in the membrane glycoprotein appears to only be present on the KR-LGG combination (of the 132 sequences with this polymorphism 131 are from Europe and 1 from North America) (Supplementary Table 2). When the other common polymorphisms (>5%) observed in the NSP2, NSP6, RNA-dependent RNA polymerase (RdRP), membrane glycoprotein and helicase are taken into account, there are at present eight main circulating strains at >5% frequency in the database all within one to three amino acid polymorphism networks (Supplementary Table 2).

Of note, our current knowledge of the global circulating strains is dependent on the ability of laboratories in different countries to deposit full genome length SARS-CoV-2 sequences and may be subject to ascertainment bias. As such, the frequencies of specific strains shown in Figure 1 may not reflect the size of the outbreak. However, the data does provide the opportunity to predict the presence of specific strains in areas given the known epidemiology within different countries and regions.

SARS-CoV-2 and Host Adaptation: Implications for global viral dynamics, pathogenesis and immunogenicity

Currently the possible functional effect(s) of the introduction of the AAACGA motif into the nucleocapsid are not known. The nucleocapsid protein is a key structural protein critical to viral transcription and assembly, suggesting that changes in this protein could either increase or decrease replicative fitness. However, it is also possible these changes could be functionally compensated by linked polymorphism in the virus and/or counterbalanced by some other host fitness benefit. However, we have not found any other polymorphism linked to the K203/R204 change to date.

Selection of viral adaptations to polymorphic host responses mediated by T cells, NK-cells, antibodies and antiviral drugs are well described for other RNA viruses such as HIV and HCV [4,11]. HIV-1 adaptations to human leucocyte antigen (HLA)-restricted T-cell responses have also been shown to be transmitted and accumulate over time [12,13]. As previously shown for SARS-CoV, T-cell responses against SARS-CoV-2 are likely to target the nucleocapsid [14]. Notably, SARS-CoV-2 R203K/G204R polymorphisms modify the predicted binding of the HLA-C*07 allele to a putative T-cell epitope containing these residues. Escape from HLA-C-restricted T-cell responses may conceivably confer a fitness advantage for SARS-CoV-2, particularly in European populations where HLA-C*07 is prevalent and carried by >40% of the population (www.allelefrequencies.net).

The replication characteristics and plasticity of small, highly mutable viruses such as HIV and HCV are distinct from SARS-CoV-2, which is significantly less variable.

However, the fact that a number of variants are established in the human pandemic indicates some capacity for successful viral diversification in response to host immunogenetic heterogeneity. The consequences of viral host adaptation for SARS-CoV-2 immunopathogenesis may also be complex. If SARS-CoV-2 T-cell responses define the specific nature of the clinical illness, such adaptation may serve to favor prolonged, perhaps asymptomatic viral shedding. This phenomena has been demonstrated in the case of mutations in the influenza genome, with deleterious public health consequences for transmission and control [15]. If the host response to SARS-CoV-2 is dependent on the efficacy or regulation of a specific antiviral cytotoxic T-cell response, adaptation may also impact individual pathways to serious complications such as acute respiratory distress syndrome and other end-organ disease. However, it is important to realize that as evidenced by HCV, HIV and other RNA viruses, any change in replicative fitness cost to the virus will not predictably translate to any attenuation of the nature of the disease in the host. As SARS-CoV-2 becomes globally established over a longer period of time, evolution towards a less immunogenic, but still highly transmissible infection could be facilitated by fixation and spread of immune-based selection events. This has been postulated to be the case for other human-adapted coronaviruses that eventually became established as seasonal "common cold" viruses.

Marked viral diversity and adaptation of other RNA viruses such as HIV and HCV to highly polymorphic host immune proteins such as HLA that restrict the immune response has been a barrier to successful vaccination to date. With the widespread global spread of SARS-CoV-2 virus across genetically diverse populations it will also be critical to elucidate the functional consequences of any newly emerging genetic

changes to guide development of antivirals and universal vaccines against conserved SARS-CoV-2 elements.

References:

1 Leslie, A. *et al.* Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *The Journal of experimental medicine* 201, 891-902, doi:10.1084/jem.20041455 (2005).

2 Leslie, A. J. *et al.* HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med* 10, 282-289, doi:10.1038/nm992 (2004).

3 Fitzmaurice, K. *et al.* Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus infection. *Gut* 60, 1563-1571, doi:10.1136/gut.2010.228403 (2011).

4 Moore, C. B. *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* 296, 1439-1443, doi:10.1126/science.1069660 (2002).

5 Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *Journal of medical virology* 92, 433-440, doi:10.1002/jmv.25682 (2020).

6 Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *Journal of virology* 84, 3134-3146, doi:10.1128/JVI.01394-09 (2010).

7 Nijhuis, M., van Maarseveen, N. M. & Boucher, C. A. HIV protease resistance and viral fitness. *Current opinion in HIV and AIDS* 2, 108-115, doi:10.1097/COH.0b013e32801682f6 (2007).

8 Rubnitz, J. & Subramani, S. The minimum amount of homology required for homologous recombination in mammalian cells. *Molecular and cellular biology* 4, 2253-2258, doi:10.1128/mcb.4.11.2253 (1984).

9 Sola, I., Moreno, J. L., Zuniga, S., Alonso, S. & Enjuanes, L. Role of nucleotides immediately flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA synthesis. *Journal of virology* 79, 2506-2516, doi:10.1128/JVI.79.4.2506-2516.2005 (2005).

10 http://www.who.int. Accessed 31st March 2020).

11 Gaudieri, S. *et al.* Evidence of viral adaptation to HLA class I-restricted immune pressure in chronic hepatitis C virus infection. *Journal of virology* 80, 11094-11104, doi:10.1128/JVI.00912-06 (2006).

12 Brumme, Z. L. *et al.* Extensive host immune adaptation in a concentrated North American HIV epidemic. *Aids* 32, 1927-1938, doi:10.1097/QAD.0000000000001912 (2018).

13 Katoh, J. *et al.* Rapid HIV-1 Disease Progression in Individuals Infected with a Virus Adapted to Its Host Population. *PloS one* 11, e0150397, doi:10.1371/journal.pone.0150397 (2016).

14 Peng, H. *et al.* Long-lived memory T lymphocyte responses against SARS coronavirus nucleocapsid protein in SARS-recovered patients. *Virology* 351, 466-475, doi:10.1016/j.virol.2006.03.036 (2006).

15 McMichael, A. J., Gotch, F. M., Noble, G. R. & Beare, P. A. Cytotoxic T-cell immunity to influenza. *The New England journal of medicine* 309, 13-17, doi:10.1056/NEJM198307073090103 (1983).

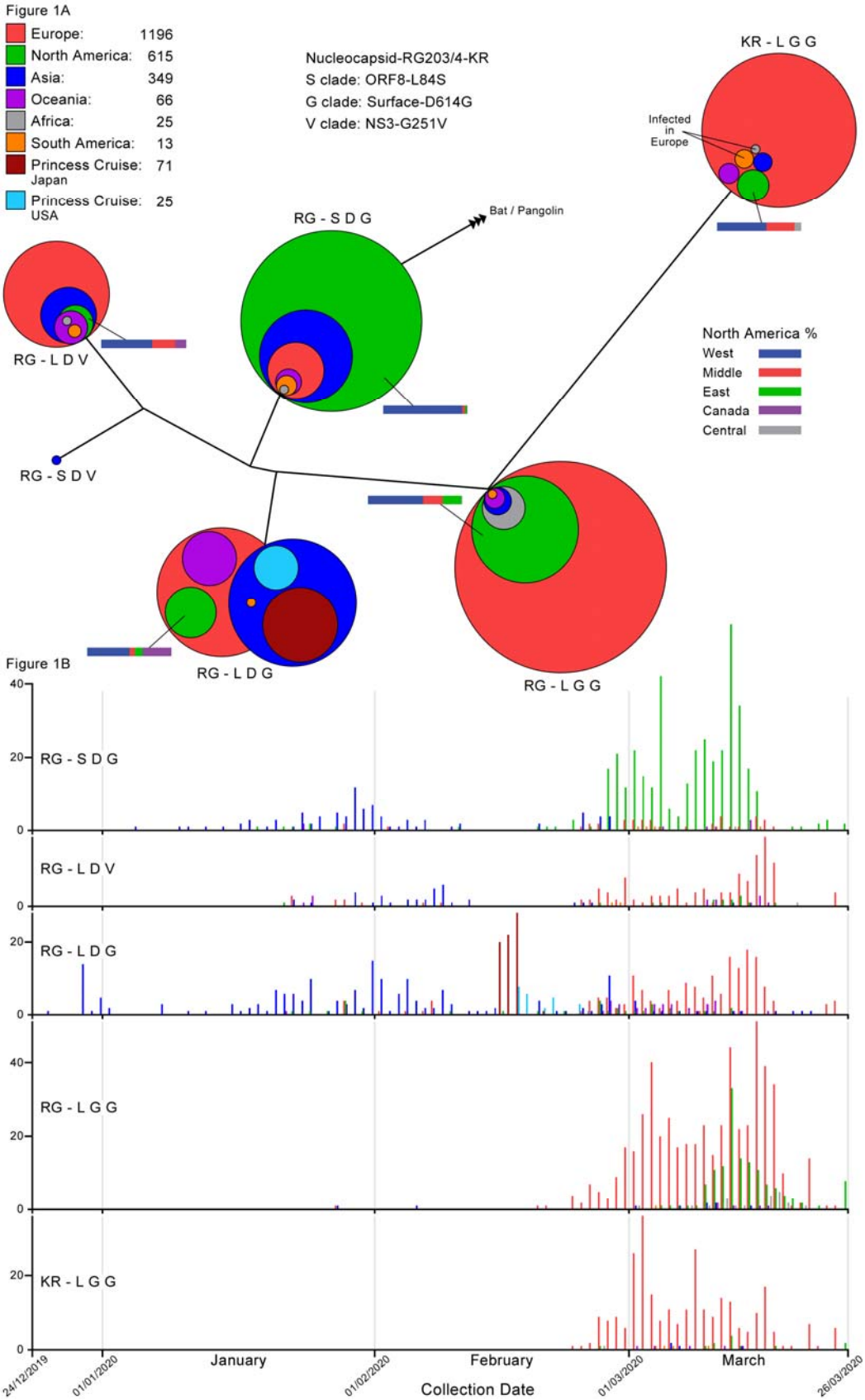

Figure 1: Emergence of new SARS-CoV-2 strains. A. The tree nodes represent the different circulating strains of SARS-CoV-2 based on the five amino acid positions indicated in the top panel. The size of the circle reflects the number of viral sequences deposited into the database and color reflects the geographical region with North America further broken down by the bar into West, East and Middle US, Canada and central America (i.e. Mexico and Panama). Princess Cruises' Diamond Princess (Japan) and Grand Princess (USA) were also separated. Note the size of the circles do not necessarily represent the size of the outbreak. **B.** The emergence of the different strains is tracked using the collection date in the metadata for the sequences with the height of the bar reflecting the number of sequences in the database with the specific strain. Color is as indicated in panel A.