Faster *de novo* mutation of SARS-CoV-2 in shipboard quarantine

Ting-Yu Yeh¹ & Gregory P. Contreras²

1. Department of Biology, Johns Hopkins University
3400 N. Charles Street, Baltimore Maryland 21218, USA

2. Columbus Center, Institute of Marine and Environmental Technology
701 East Pratt Street, Baltimore, Maryland 21202, USA

Correspondence to Ting-Yu Yeh (email: yehty@auxergen.com)

(Submitted: 5 April 2020 – Published online: 6 April 2020)

**DISCLAIMER**

This paper was submitted to the Bulletin of the World Health Organization and was posted to the COVID-19 open site, according to the protocol for public health emergencies for international concern as described in Vasee Moorthy et al. ([http://dx.doi.org/10.2471/BLT.20.251561](http://dx.doi.org/10.2471/BLT.20.251561)).

The information herein is available for unrestricted use, distribution and reproduction in any medium, provided that the original work is properly cited as indicated by the Creative Commons Attribution 3.0 Intergovernmental Organizations licence (CC BY IGO 3.0).

**RECOMMENDED CITATION**

Yeh TY & Contreras GP. Faster *de novo* mutation of SARS-CoV-2 in shipboard quarantine. [Preprint]. *Bull World Health Organ.* E-pub: 6 April 2020. doi: [http://dx.doi.org/10.2471/BLT.20.255752](http://dx.doi.org/10.2471/BLT.20.255752)
Abstract

Objective To examine transmission and evolution of SARS-CoV-2 in shipboard quarantine.

Methods 28 specimens of COVID-19 patients from the cruise with a 3 week quarantine period were collected in USA and Japan from February 10 to February 25, 2020. The full genome sequences of SARS-CoV-2 are available on GISAID. Viral transmission and evolution dynamics were analyzed using computational tools of phylogenetics, natural selection pressure, and genetic linkage.

Findings The SARS-CoV-2 outbreak in the cruise appears to have originated from either a single primary case with viral sequence identical to WIV04 isolates, or simultaneously with the other primary case with G11083T mutation. 24 new viral mutations are present in 64.2% (18/28) patient samples in the cruise, and the virus has already evolved into at least five subgroups. Viral transmission occurred more commonly between these subgroups. Accelerated mutation and increased positive selection of SARA-CoV-2 are statistically significant during the quarantine (Tajima’s $D$: -2.03, $p<0.01$; Fu and Li’s $D$: -2.66, $p<0.01$; and Zeng’s $E$: -2.37, $p<0.01$). Linage disequilibrium analysis confirmed that RNA recombination with G11083T mutation also contributes to the increase of mutations among the viral progeny.

Conclusion Despite the fact that the quarantine reduced infections, this study shows faster mutation of SARS-CoV-2 shipboard. This is also the first evidence that de novo RNA recombination accelerates SARA-CoV-2 evolution under positive selection pressure. Therefore, health authorities need to revise quarantine protocols in close-quarter environments.
Introduction

On December 31 of 2019, Chinese authorities alerted the World Health Organization (WHO) of an outbreak of a novel coronavirus causing severe illness with pneumonia-like symptoms, which was later named SARS-CoV-2.\(^1\) The outbreak was declared a Public Health Emergency of International Concern on 30 January 2020, and WHO declared COVID-19 a pandemic on March 11, 2020\(^2\). As of April 4, 2020, more than 1.2 million COVID-19 cases have been confirmed worldwide. The virus has killed over 65,000 people, although many asymptomatic cases also likely exist without diagnosis.

Immediately after the epidemic of COVID-19 began, scientists from various countries sequenced the SARS-CoV-2 genome and made the data available to researchers in GISAID worldwide. The viral genomic sequences have indicated that the number of COVID-19 cases increased due to human to human transmission likely after a single introduction into the human population\(^3\). Many scientists used this sequencing data to explore the origins and evolution of SARS-CoV-2, SARS-CoV and related bat coronavirus RaILT3. Importantly no evidence was found to support the hypothesis that SARS-CoV-2 was engineered\(^3\).

SARS-CoV-2 evolves quickly through random nucleotide substitutions. Since, only five mutations varied in the nucleotide sequences of the SARS-CoV-2 isolates also suggest short infection history in humans\(^4\). So far more than 131 mutations SARS-CoV-2 have been found, including mutations in 149 sites across the 103 sequenced strains. Ancestral states for 43 synonymous, 83 non-synonymous, and two stop-gain mutations were unambiguously inferred\(^5\). Tang et al. categorized SARS-CoV-2 genomes into two major types based on the nucleotide 28,144 as T or C (L- or S-type, respectively) after
evolution. Over the length of nearly 30,000-base-pair genome, SARS-CoV-2 accumulates an average of about one to two mutations per month⁵.

On January 25, 2020, a passenger disembarked a cruise ship in Hong Kong and on February 1 they tested positive for SARS-CoV-2. The ship docked in Yokohama, Japan, on February 3 for quarantine and isolation. Among 3,711 cruise passengers and crew, 712 (19.2%) had been tested positive for SARS-CoV-2, and 331 (46.5%) were asymptomatic at the time of testing. Among 381 symptomatic patients, 37 (9.7%) required intensive care, and nine (1.3%) were deceased⁶,⁷. As of March 13, among 428 U.S. passengers and crew, 107 (25.0%) had positive test results; 11 U.S. passengers remained hospitalized in Japan, including seven in serious condition⁶.

However, whether the shipboard quarantine creates favorable environment for viral mutation remains unclear. Moreover, the evolution process of SARS-CoV-2 has been not closely observed. Yet, cruise ships present an ideal static population to measure the viral phylodynamics from the COVID-19 outbreak, providing the first insight into the de novo evolution of SARS-CoV-2 in a closed population.

Methods

Data resources

A total of 28 viral nasopharyngeal or oropharyngeal swab specimens isolated from the cruise patients from February 10 to February 25 were collected (Table 1). Viral genomes were Sanger sequenced with Nanopore MiniON and Illumina MinSeq. Sequence assembly was completed using minimap software, version 2.17 (Minion), Burrows-Wheeler-Aligner (BWA), version 0.7.17 (Illumina), and Sequencher software, version 5.4.6 (Sanger). The viral sequence of deceased case 27 was sequenced using NextSeq500
(Illumina) at 1500 X mean coverage and assembled with A5-miniseq software, version 20140604. Viral sequences and sequencing methods are deposited and available in GISAID (Global Initiative on Sharing All Influenza Data; https://www.gisaid.org/) and GenBank (https://www.ncbi.nlm.nih.gov/genbank) at (1) Pathology Discovery, Respiratory Viruses Branch, Division of Virus Diseases, Centers for Disease Control (CDC), Atlanta, USA (case 1 to 26, case 20 was missing in the record); (2) Department of Pathology of Toshima Hospital, and Pathogen Genomics Center of National Institute of Infectious Diseases, Tokyo, Japan (case 27); and (3) Department of Microbiology, Kanagawa Prefectural Institute of Public Health, Kanagawa, Japan (case 28 and 29). Sequences and annotations of the reference genomes of SARS-CoV-2 isolates PBCAMS-WH-04, WIV04, Hu-1, and WHU01 (access number MT019532, MN996528, NC045512, and MN98868)\textsuperscript{4} were downloaded from GenBank or GISAID.

**Statistical and phylogenetics analyses**

FASTA files of viral sequences were aligned using MAFFT 7 (https://mafft.cbrc.jp/alignment/server/index.html).\textsuperscript{8} Phylogenetic relationships between viral sequences were analyzed using the neighbor-joining method and Jukes-Cantor substitution model with setting bootstrap resampling number as 5. The rectangular phylogenetic tree was generated using Archaeopteryx with Java plugin.\textsuperscript{9} For radial phylogenetic tree, the tree file was first exported as Newick format by MAFFT, the cladogram was transformed and displayed by FigTree software, version 1.4.2 (Andrew Rambaut, tree.bio.ed.ac.uk/software/figtree). Rooting an unrooted tree was done by introducing an outgroup virus, which is bat SARS-like coronavirus WIV16 (accession number KT444582).
To determine whether virus genome polymorphisms, undergo neutral or non-neutral evolution, MEGA7 software was used to calculate Tajima’s $D$ test of neutrality\textsuperscript{11}, which compares the number of segregating sites per site with the nucleotide diversity (the mean pairwise difference between sequences). Fu and Li’s $D$ (comparison of the number of derived singleton mutations and the mean pairwise) and Zeng’s $E$ values were measured using online PopSc calculator (popsc.chenshiyi.com)\textsuperscript{12-14}.

To determine the linkage disequilibrium (LD) of SARS-Cov-2 genomes, 148 SARS-Cov-2 genomic sequences from GISAID was first converted using SNP_tools plug-in in Microsoft Excel\textsuperscript{15}. The values of $D'$ (The coefficient of linkage disequilibrium/ $D_{max}$), LOD (log of the odds of there being LD between two loci), and $r^2$ (squared coefficient of correlation) were measured and plotted by HaploView software, version 4.1.\textsuperscript{16} In the absence of evolutionary forces or natural selection, the $D'$ converges to zero along the time axis at a rate depending on the magnitude of the recombination rate between the two loci. Chi-square test was used to examine if LD is statistically significant. For RNA recombination, confidence bounds were plotted by HaploView, based on the default algorithm proposed by Gabriel et al.\textsuperscript{17}, for which 95% confidence bounds on $D'$ are generated. Pairs are thought to be in “strong LD” if the one-sided upper 95% confidence bound is $> 0.98$ (that is, consistent with no historical recombination) and the lower bound is above 0.7. Conversely, “strong evidence for historical recombination” is defined if pairs for which the upper confidence bound on $D'$ is less than 0.9. A "spine" of strong LD running from one marker to another along the legs of the triangle in the LD chart was searched to determine the block\textsuperscript{16}. Haplotypes are estimated using an accelerated expectation-maximization algorithm similar to the partition/ligation method\textsuperscript{18}. This
creates highly accurate population frequency estimates of the phased haplotypes based on the maximum likelihood as determined from the unphased input.

**Results**

**New viral mutations in the cruise**

There are 28 specimens with viral sequences available so far, including 25 samples from USA and 3 samples from Japan. The detail information of virus name, accession ID, specimen source, collection date, and mutations from the cruise cases are listed in Table 1. Genetic variations of viral sequences are present in 71.4% (20/28) samples. The sequences of case 3, 5, 9, 13, 15, 16, 19, 22 are completely identical to the Wuhan isolates PBCAMS-WH-04, WIV04, Hu-1, and WHU01 (access number MT019532, MN996528, NC045512, and MN98868). Except for the G11083T mutation, a total of 24 new substitution mutations were identified (Table 1 and Figure 1A). Night synonymous mutations (C1385T, T9157C, C10036T, C10507T, G11410A, C11956T, G28378T, C29230T, C29635T) are present in the viral protein coding region. Eleven mutations (C3099T, G3259T, C3738T, A5845T, C6636T, C9474T, C11750T, C22033A, G22104T, C28409T, G28916A) lead to nonsynonymous substitutions in viral proteins. Four mutations exist in the non-coding region of SARS-CoV-2 genome, including C254T (in 5’ untranslated region), C26326T (in intergenic region between S protein orf3a), G29736T and G29751C (in 3’ untranslated region). 57.1% (16/28 cases) samples have nonsynonymous substitutions: case 1, 2, 7, 8, 10, 17, 25, 27, 28, 29 have one, case 8, 14, 18, 23, 24, 26 have two. These non-silent mutations affected three proteins: ORF1ab (T945I, Q998H, P1158L, K1860N, T2124I, A3070V, L3828F), spike protein (G181V and F156) and nucleocapsid protein (P46S and G214S).
Two mutations have previously been identified in other SARS-Cov-2 isolates. Case 18 has 15 nucleotide deletion between nucleotide 508 to 522, such deletions were also detected in USA-CA6 case (access number MT044258). However, the possibility that case 18 contributes to the new mutation in the cruise can be ruled out because this deletion is absent in the other cruise specimens. The other nonsynonymous mutation (ORF1ab L3606F), G11083T, was also found in USA-AZ1 case earlier on January 22, (accession number EPI_ISL_406223). Later G11083T mutation was also present in WA3-UW1 (Washington, USA, February 27), NY-NYUMC1 (New York City, USA, March 3), and UPHL-01 (Utah Public Health Laboratory, Salt Lake City, March 13) isolates (Accession number EPI_ISL_413025, EPI_ISL_414639, and EPI_ISL_415539).

**Viral origin and transmission**

To investigate the infectious origin and virus transmission process in the cruise, the viral phylodynamics was analyzed using neighbor joining method with or without introducing an outgroup virus (bat coronavirus SARS-WIV16) to allow for genome sequences to be rerooted\(^8,9,\)\(^{19}\). Both phylogenetic trees are shown in Figure 1. The data indicates that 50% (14/28) of specimens contain the virus that has already evolved into several subgroups during the shipboard quarantine: case 1/24 (C3099T and G28378T), case 25/26/29 (G11083T), case 6/7 (G11410A and C26326T), case 4/12/27/28 (C29635T), case 2/21/23 (G29736T) (Figure 1A and 1B). Due to G11083T mutation, the clustering of taxa on viral phylogenies is obvious with spatially structured host population between these subgroups (50%, 14/28 cases) (Figure 1A, mutations labeled in colors). This result also indicates that viral transmission occurred more commonly between these subgroups.
So far, it remains uncertain that this single G11083T substitution spontaneously occurred during the quarantine or the patients had been infected outside. Nevertheless, all of the cruise viral samples shared more sequence identity with the WIV4 sequence than with other 143 SARS-CoV-2 isolates in the database. It strongly suggests that these 24 new mutations were generated in the cruise *de novo* as community infection rather than deriving from multiple geographic origins. This phylogenetic analysis reveals two possibilities of the infectious origin: the SARS-CoV-2 infection (except case 18) in the cruise was originated from (1) a single primary case with WIV04 sequence, then all mutations (including G11083T) occurred *de novo* during the quarantine, or (2) two simultaneously primary cases, one with WIV04 sequence and the other one with G11083T substitution.

**Natural selection of viral mutations**

Three mutations were present in case 4, 7, 14, and 21, and four in case 24. Case 26 has total 6 nucleotide mutations (Figure 1 and Table 1). Since SARS-CoV-2 accumulates an average of about one to two mutations per month\(^5\,20\), the viral mutation rate appears to be much faster than normal with regards to the 3 week quarantine period. This observation leads to the hypothesis that the virus mutation accelerates under selection pressure as opposed to neutral evolution (random) onboard the cruise ship. Because the differences among conspecific SARS-CoV-2 sequences in this study represent transient polymorphisms rather than fixation events along independent lineages over long time scales, the popular tool like dN/dS (\(\omega\)) ratio fail to measure selection pressure accurately\(^21,22\) (see Discussion). To test this hypothesis, Tajima \(D\) statistics was calculated by MEGA software to test neutrality of DNA polymorphisms because it is little affected by the sample size, number of segregation sites, or recombination events\(^11,23\). The
Tajima’s $D$ value in the cruise is -2.03 ($p<0.01$), compared to 6.75 among total 140 full length genomic sequences of SARS-CoV-2 isolates currently available. This result indicates that, while other SARS-CoV-2 isolates face balancing selection, the virus in the cruise has evolved under two possible but not exclusive forces generated during the quarantine procedure: (1) purifying or positive selection, and (2) population growth after a recent bottleneck.

These two possibilities were further verified in other neutrality tests, such as Fu and Li’s $D$, and Zeng’s $E$ test$^{12,13}$. Possibility (1) is evident due to the negative $D$ value ($D= -2.66$, $p<0.01$) calculate by Fu and Li’s test and suggests that the quarantine procedure provided certain evolutionary pressure to generate an excess of singletons. This conclusion is also supported by the fact that the 39.2% (11/28) of cases contain 15 new mutations (mutations labeled in black in Figure 1A) that has only occurred once so far. Possibility (2) is very likely true because Zeng’s $E$ value ($E= -2.37$, $p<0.01$) is also negative. It is reasonable to assume that viral sequence polymorphisms of SARS-CoV-2 are at an initial stage of evolution rather than the fixation stage in such a short incubation time. The advantage of Zeng’s $E$ value is that this test is not affected by positive selection at various evolution stages but denotes very powerful evidence for the population growth theory$^{13}$. Taken together, it is reasonable to conclude that SARS-CoV-2 evolves faster during the shipboard quarantine, and the viral evolution is positively correlated to the increase of the selection pressure.

**RNA recombination**

Case 24 and 23 have the identical G11083T mutation as cases 25, 26 and 29 despite the fact that they do not belong to the same group. The similar phenomena is also observed in Japan patients in the cruise: cases 27 and 28 have both C29635T and
G11083T mutations. By assuming the substitution rate of $0.92 \times 10^{-3}$/site/year as estimated in SARS-CoV-2\textsuperscript{20}, it is very unlikely the virus of case 24, 23, 27, and 28 all coincidently mutated the identical nucleotide G11083 into T within 3 weeks. Many coronaviruses are simultaneously maintained in nature, which allows for genetic recombination and results in new virus strains\textsuperscript{25}. One rational hypothesis is that RNA recombination occurs in these cases to gain G11083T mutation. However, RNA recombination in SARS-CoV-2 have not been reported after analyzing 113 isolates of SARS-CoV-2 so far\textsuperscript{5}.

To determine whether case 23, 24, 27 and 28 obtain the same G11083T mutation quickly via RNA recombination, HaploView software was used to analyze and visualize the patterns of linkage disequilibrium (LD) between variants with minor alleles of two SARS-CoV-2 strains (Fig. 2). Among samples with the linkage disequilibrium measure \((D')\) value equal to 1 (red squares in Figure 2A), four pairs of mutations show high LOD (> 3.0) and \(r^2\) value, including C3099/G28378 (LOD = 4.6, \(r^2 =1\)), G11410/C26326 (LOD =3.77, \(r^2 =0.66\)), G29736/29751 (LOD =3.77, \(r^2 =0.66\)), C254/G29736 (LOD =3.77, \(r^2 =0.66\)), C254/G29751 (LOD = 4.6, \(r^2 =1\)) (black squares in Figure 2B and 2C). This data indicates that linkages between double mutations in case 1/24, 6/7, 21/23 are statistically significant (\(p < 0.001\) in chi-square test, df =1).

Positive RNA recombination was determined by 95% confidence bounds on \(D'\), and "strong recombination" was defined if pairs for which the upper confidence bound on \(D'\) is less than 0.9\textsuperscript{17}. The upper confidence bounds on \(D'\) of C254/G11083, G29726/G11083, G29751/G11083 (case 23), C3099/G11083, G28378/G11083 (case 24), C29635/G11083 (case 27, 28) are 0.86, 0.73, 0.86, 0.86, 0.86, and 0.80, respectively (white squares, Figure 2D). This result strongly supports the hypothesis that G11083T mutation has been
transmitted via RNA recombination in case 23, 24, 27 and 28. It is worthwhile to mention the UPHL-01 case, which has G11083T mutation and unique C26326T is only found in cases 6 and 7 on the cruise but not in other isolates in the world (Figure 1A). The upper confidence bounds on $D'$ of C26326/G11083 is 0.73 (Figure 3D), suggesting that RNA recombination event may have also occurred in UPHL-01 recently.

The haplotype block was obtained by the solid spine of LD method\textsuperscript{16}, each haplotype in a block with its population frequency and connections between block 1 (red) and block 2 (blue) was shown in Figure 3D. The value of multiallelic $D'$ is 65%, which represents the level of recombination between the two blocks. Taken together, these data provide the first evidence of RNA recombination contributes to SARS-CoV-2 mutations, which occurred during shipboard quarantine.

**Novel Viral mutations**

The viral populations in the cruise also generated 11 new variants in the viral proteins, including ORF1ab (NSP3, NSP4, NSP6), and in the viral structural spike and nucleocapsid protein (Table 1). There are 5 missense mutations found in NSP3 protein of ORF1ab: C3099T and G3259T causing the replacement of threonine 945 with isoleucine (T945I, case 1) and glutamine 998 with histidine (Q998H, case 10), respectively. C3738T (case 7) results in proline 1158 changing into leucine (P1158L) in the macro domain, and lysine 1860 mutates to asparagine (K1860N) in papain-like protease (PLpro) due to A5845T (case 17). C6636T causes threonine 2124 converting to isoleucine (T2124I, case 14) in the group 2 marker (G2M) domain. In NSP4 protein, alanine to valine mutation in ORF1ab amino acid 3070 (A3070V) was led by C9474T (case 8). Replacement of leucine 3828 with phenylalanine (L3828F) occurs in NSP6 in case 14 (C11750T). In case 18 (C22033A) and 23 (G22104T), phenylalanine 156 and glycine 181 of coronavirus
spike protein were replaced by leucine and valine (F156L and G181V), respectively. Both proline 46 and glycine 214 change to serine (P46S in RNA binding domain and G214S in arginine-serine domain) in the nucleocapsid protein in case 2 (C28409T) and 26 (G28916A).

It is worthy to note that new viral mutations (G29736T, G29751T) in case 2, 21, 23 are located at the stem loop-II (s2m) motif, an extremely conserved RNA element in the 3’ untranslated region including more than 30 coronaviruses such as SARS26 (Figure 3A). Based on the 3D crystal structure of SARS-CoV s2m RNA27 (Figure 3B), G29736 and G29751 are equal to essential nucleotides “G13” and “G28” in SARS-CoV s2m. Both nucleotides are essential to form the critical viral RNA structure that mimics ribosomal binding27. Taken together, SARS-Cov-2 in the population during the shipboard quarantine already generated new mutations that may affect viral protein or RNA functions.
Discussion

The shipboard quarantine provides a near-perfect incubation environment to observe the SARS-CoV-2 transmission and adaptation independently from other infectious resources\(^6,7\). Thus, it enables researchers to observe viral transmission of SARS-CoV-2 with very little noise. Due to the impact transmission dynamics and selection have on viral genetic variation, viral phylogenies can therefore be used to investigate important epidemiological and evolutionary processes, such as epidemic spread and spatio-temporal dynamics including metapopulation dynamics, zoonotic transmission, tissue tropism, and antigenic drift.\(^28\) This study reports the viral phylodynamics of SARS-CoV-2 from the patients in the shipboard quarantine for 3 weeks period in this February, 2020. The transmission started from either one or two primary cases with WIV04 sequence and / or G29751T mutation, then quickly separated into at least 5 subgroups based on new mutations. Accelerated mutation and increased positive selection of SARA-CoV-2 are evident during the quarantine. Furthermore, RNA recombination also contributes to the increase of mutation rate. This experience should be carefully considered in formulation of new management protocols with respect to a SARA-CoV-2 outbreak in any relatively close quarters, such as shipboard, in submarines, dormitories, hospitals, etc.

Russell et al. estimated case and infection fatality ratios (CFR, IFR) for COVID-19 on the Diamond Princess ship as 2.3% (0.75%-5.3%) and 1.2% (0.38-2.7%)\(^29-31\). Using mathematical modeling, Mizumoto and Chowell’s estimate of the mean reproduction number in the confined setting reached values as high as \(\sim 11\), which is higher than mean estimates reported from community-level transmission dynamics in China and Singapore (approximate range: 1.1-7). Their findings suggest that \(R_t\) substantially decreased compared to values during the early phase after the Japanese government implemented an
enhanced quarantine control. Most recent estimates of $R_t$ reached values largely below the epidemic threshold, indicating that a secondary outbreak of SARS-CoV-2 was unlikely to occur aboard the Diamond Princess Ship\textsuperscript{32}. Despite the fact that the quarantine averted a lot of infections, the phylogenetics analysis clearly detects viral transmission and RNA recombination occurred more commonly between these subgroups. Our data fit in the coalescent model, which the two lineages are traced back to two genetically distinct variants present at transmission\textsuperscript{24}. However, it can not be ruled out that evolutionary processes also potentially interacted to shape the viral phylogenies. While spatial structure is the most general population structure in phylodynamic analyses as well as in the case of the cruise, SARS-CoV-2 may also have a non-random admixture of attributes including the age, race, and risk behavior, etc.\textsuperscript{28} Because viral transmission can preferentially occur between patients sharing any of these attributes, the real reason(s) for viral transmission between these subgroups require(s) further study. Caution needs to be taken to evaluate whether quarantine in close-quarters also promote virus to mutate more rapidly.

Although the information of the mutation rate is limited in this study due to the relatively small sample size, it is obvious that the virus mutated faster onboard the cruise ship, and the selection pressure is not random based on computational and statistic tools. Despite the simplicity and robustness of $dN/dS$ ratio to measure selection pressure of adaptive evolution, this popular method was not employed here owing to the following reasons: First, $dN/dS$ ratio is defined to describe the relative rate of selected versus neutral fixation events over long time-scales, not transient polymorphisms over short time-scales. Second, it has been shown that $dN/dS$ value in a single population does not follow a monotonic function in proportion to positive selection pressure like diverged
sequences. Therefore, Tajima’s $D$ test is used because its power is not affected by RNA recombination and various time scales, nor does this test require the outgroup sequence or large sample size. Fu and Li’ $D$ test is similar but more sensitive when there is an excess of singletons, which are also present in the cruise populations. Zeng’s $E$ test has the advantage because it is almost absent of any power at this early stage of virus transmission, the negative value is mainly contributed by the population growth. These neutrality statistics all lead to the conclusions that, on the cruise, the virus evolved under strong positive selection or maybe in the process of selective sweeps, which generates beneficial mutation for SARS-CoV-2 to increase its frequency and reach fixation quickly. It has been suggested that human intervention may have placed more severe selective pressure on the L type of SARS-CoV-2, which might be more aggressive and spread more quickly. Caution needs to be taken to evaluate whether quarantine procedure in a closed community actually provides the incubation for virus to mutate or adapt to human hosts more rapidly.

Although RNA recombination in SARS-CoV-2 was not found previously, this study provides the first evidence that RNA recombination occurred de novo in SARS-CoV-2 genome via LD analysis. Within 3 weeks, case 23, 24, 27 and 28 have gained the same G11083T mutation, suggesting RNA recombination also participates in viral evolution in combination with positive pressure. So far there is no evidence if UPHL-01 case was linked to cruise cases during the viral transmission process or if it acted independently, but RNA recombination of G11803T also occurred in UPHL-01 case, suggesting similar recombination mechanism may also exist in this case. Because G11083T mutation was also later present in other cases WA3-UW1, NY-NYUMC1 in addition to UPHL-01, it
requires attention whether G11083T may increase the fitness of the carrier as a benefit allele in the future.

The 11 missense mutations that affect amino acid in ORF1ab, spike and nucleocapsid proteins were identified in this study. It has been shown that recombination of SARS coronavirus may affect nucleocapsid protein multimerization\(^3^3\). In the 3’ untranslated region of viral genomic RNA, s2m is a very well-conserved RNA motif among beta coronaviruses\(^2^5\). A possible mechanism has been proposed for SARS viral RNA hijacking of host protein synthesis, both based upon observed s2m RNA macromolecular mimicry of a relevant ribosomal RNA fold\(^2^6\). Unique single G29742A or G29742U substitution in s2m motif has been reported in SARA-CoV-2 isolates in Australia\(^3^4\). This study reinforces the idea that the conserved s2m domain becomes the hotspot of SARA-CoV-2 mutation. G29736, which is equal to “G13” in SARS-CoV s2m, forms a base triple with A38 and C39 in a seven-nucleotide asymmetric bubble (Figure 3B, green). G29751 acts as “G28” of SARS-CoV s2m, which participates in formation of an essential RNA base quartet composed of two G-C pairs (G19, C20, G28, C31)\(^2^6\). Such mutations may disrupt RNA structure and thereby alter the viral viability or infectivity dramatically. The possibility that these mutations in viral protein or regulatory RNA regions may enhance adaptation or attenuate virus replication or virulence\(^3^5\) requires further investigation.
References

1. Novel Coronavirus – China. Geneva: World Health Organization; 2020. Available from https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/

2. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Geneva: World Health Organization; 2020. Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

3. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Med. 2020 March 17; https://doi.org/10.1038/s41591-020-0820-9

4. Lu R, Zhao X, Li J, Niu P, Yang B, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020 Feb 22;395(10224):565-574. doi: 10.1016/S0140-6736(20)30251-8. PMID: 32007145

5. Tang X, Wu C, Li X, Song Y, Yao X et al. On the origin and continuing evolution of SARS-CoV-2. Nat Sci Rev, 2020 March 03; https://doi.org/10.1093/nsr/nwaa036

6. Moriarty LF, Plucinski MM, Marston BJ, Kurbatova EV, Knust B, et al. Public health responses to COVID-19 outbreaks on cruise ships - Worldwide, February-March 2020. MMWR Morb Mortal Wkly Rep. 2020 Mar 27;69(12):347-352. doi: 10.15585/mmwr.mm6912e3. PMID: 32214086

7. Kakimoto K, Kamiya H, Yamagishi T, Matsui T, Suzuki M, Wakita T. Initial Investigation of Transmission of COVID-19 among crew members during quarantine of a cruise ship - Yokohama, Japan, February 2020. MMWR Morb
8. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-80. doi: 10.1093/molbev/mst010. PMID: 23329690

9. Kuraku S, Zmasek CM, Nishimura O, Katoh K. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W22-8. doi: 10.1093/nar/gkt389. PMID: 23677614

10. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis Version 7.0 for bigger datasets. Mol Biol Evol. 2016 Jul;33(7):1870-4. doi: 10.1093/molbev/msw054. PMID: 27004904

11. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989 Nov;123(3):585-95. PMID: 2513255

12. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993 Mar;133(3):693-709. PMID: 8454210

13. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006 Nov;174(3):1431-9. https://doi.org/10.1534/genetics.106.061432. PMID: 16951063

14. Chen SY, Deng F, Huang Y, Li 1, Liu L, et al. PopSc: Computing Toolkit for Basic statistics of molecular population genetics simultaneously implemented in web-based calculator, Python and R. PLoS One. 2016 Oct 28;11(10):e0165434. doi: 10.1371/journal.pone.0165434. PMID: 27792763
15. Chen B, Wilkening S, Drechsel M, Hemminki K. SNP_tools: A compact tool package for analysis and conversion of genotype data for MS-Excel. BMC Res Notes. 2009 Oct 23;2:214. doi: 10.1186/1756-0500-2-214. PMID: 19852806

16. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005 Jan 15;21(2):263-5. PMID: 15297300

17. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al., The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9. DOI: 10.1126/science.1069424 PMID: 12029063

18. Qin ZS, Niu T, Liu JS. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet. 2002 Nov;71(5):1242-7. doi: 10.1086/344207 PMID: 12452179

19. Poon LL, Chu DK, Chan KH, Wong OK, Ellis TM, et al. Identification of a novel coronavirus in bats. J Virol. 2005 Feb;79(4):2001-9. DOI: 10.1128/JVI.79.4.2001-2009.2005 PMID: 15681402

20. Fang B, Liu L, Yu X, Li X, Ye G, et al. Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2). bioRxiv 2020 Mar 11; doi: https://doi.org/10.1101/2020.03.04.9766622

21. Yang W, Bielawski JP, Yang Z. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J Mol Evol. 2003 Aug;57(2):212-21. DOI: 10.1007/s00239-003-2467-9 PMID: 14562964

22. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008 Dec;4(12):e1000304. doi: 10.1371/journal.pgen.1000304. PMID: 19081788
23. Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics. 2008 May;179(1):555-67. doi: 10.1534/genetics.107.083006. PMID: 18493071

24. McCrone JT, Lauring AS. Genetic bottlenecks in intraspecies virus transmission. Curr Opin Virol. 2018 Feb;28:20-25. doi: 10.1016/j.covalro.2017.10.008. PMID: 29107838

25. Su S, Wong G, Shi W, Liu J, Lai ACK, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 2016 Jun;24(6):490-502. doi: 10.1016/j.tim.2016.03.003. PMID: 27012512

26. Tengs T, Jonassen CM. Distribution and evolutionary history of the mobile genetic element s2m in coronaviruses. Diseases. 2016 Jul 28;4(3). pii: E27. doi: 10.3390/diseases4030027. PMID: 28933407

27. Robertson MP, Igel H, Baertsch R, Haussler D, Ares M Jr, Scott WG. The structure of a rigorously conserved RNA element within the SARS virus genome. PLoS Biol. 2005 Jan;3(1):e5. PMID: 15630477 DOI:10.1371/journal.pbio.0030005

28. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9(3):e1002947. doi: 10.1371/journal.pcbi.1002947. PMID: 23555203

29. Russell TW, Hellewell J, Jarvis JJ, van-Zandvoort K, Abbott, S, et al. Estimating the infection and case fatality ratio for COVID-19 using age-adjusted data from the outbreak on the Diamond Princess cruise ship. Preprint at medRxiv. https://doi.org/10.1101/2020.03.05.20031773 (2020).
30. National Institute of Infectious Diseases. Field Briefing: Diamond Princess COVID-19 Cases. https://www.niid.go.jp/niid/en/2019-ncov-e/9407-covid-dp-fe-01.html (accessed 3 Mar 2020).

31. National Institute of Infectious Diseases. Field Briefing: Diamond Princess COVID-19 Cases. https://www.niid.go.jp/niid/en/2019-ncov-e/9417-covid-dp-fe-02.html (accessed 3 Mar 2020).

32. Mizumoto K, Chowell G. Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. Infect Dis Model. 2020 Feb 29;5:264-270. doi: 10.1016/j.idm.2020.02.003. PMID: 32190785

33. Luo H, Ye F, Chen K, Shen X, Jiang H. SR-rich motif plays a pivotal role in recombinant SARS coronavirus nucleocapsid protein multimerization. Biochemistry. 2005 Nov 22;44(46):15351-8. DOI: 10.1021/bi051122c PMID: 16285739

34. Yeh TY. Emerging viral mutants in Australia SARS-CoV-2 (COVID-19) genome. (under review)

35. Muth D, Corman VM, Roth H, Binger 3, Dijkman R. et al. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Sci Rep. 2018 Oct 11;8(1):15177. doi: 10.1038/s41598-018-33487-8. PMID: 30310104

Author contribution: T.-Y. Yeh and G.P. Contreras are responsible for literature search, figures, data analysis, data interpretation and writing.

Competing interests: not declared.

Funding: not declared.
Figure 1 (A) Phylogenetic tree of the viral sequences from the cruise was generated in MAFFT using the neighbor-joining method. Bootstrap values for 5 replicates are indicated. Same mutations are shown in the same colors. (B) Alignments were generated using MAFFT, and the phylogenetic tree was visualized using FigTree. Rooting was done by introducing bat SARS-like coronavirus WIV16 as an outgroup virus.
Figure 2. Haplotype block organization of SARS-CoV-2 mutations in the cruises. (A) The standard color scheme of HaploView was used to display the strength of LD. The number near slashes at the top of the image shows the coordinate of sites in the genome. Color in the square is given by standard ($D'/LOD$), and the number in square is $D'$ value. The squares are shown as bright red ($LOD>2, D'=1$), blue ($LOD<2, D'=1$), and white ($LOD<2, D'<1$). (B) LD plot with alternate $D'/LOD$ color scheme. Squares with low LOD and high $D'$ are in shades of pink. Squares with high LOD and $D'$ are in black, which correspond to the double mutations in specific cases (color asterisks). (C) LD plot with $r^2$ values. Squares with $r^2=1$ value are in black. (D) LD Plot with confidence bounds color scheme. "Strong evidence recombination" is defined if pairs for which the upper confidence bound on $D'$ is less than 0.9 (white squares). Recombination occurs in the cases 23, 24, 27, 28, and UPHL-1 (asterisks). (E) Numbers next to each haplotype block are haplotype frequencies. Bold lines joining haplotypes from each block represent combined haplotypes with frequencies >10%, and thin lines are for frequencies >1%. In the crossing areas between haplotype blocks, a value of multi-allelic $D'$ is shown to represent the level of recombination between blocks.
Figure 3. (A) Mutations in the SARS-CoV-2 s2m RNA motif in case 2, 21, and 23. Conventional RNA helical base-pairing are indicated in italics. Sequence complements are indicated using color-coded brackets. “G13” and “G28” mutation (according to SARS-CoV$^{27}$, arrowhead) are shown with purple and orange color, respectively. (B) Schematic representation of the s2m RNA secondary structure of SARS virus, with tertiary structural interactions indicated as long range contacts$^{27}$. 
Table 1. The information of the cruise cases from February 10 to February 25.

| Virus name | Case | Accession ID | Specimen source      | Collection date | Mutations                                                                 |
|------------|------|--------------|----------------------|----------------|---------------------------------------------------------------------------|
| hCoV-19/USA/CruiseA-1/2020 | 1    | EPI_ISL_413606 | Nasopharyngeal swab  | 2020-02-17     | C3099T (Orf1a T945I), G28378T                                            |
| hCoV-19/USA/CruiseA-2/2020 | 2    | EPI_ISL_413607 | Nasopharyngeal swab  | 2020-02-18     | C28409T (N protein P46S), G29736T                                        |
| hCoV-19/USA/CruiseA-3/2020 | 3    | EPI_ISL_413608 | Nasopharyngeal swab  | 2020-02-18     |                                                                           |
| hCoV-19/USA/CruiseA-4/2020 | 4    | EPI_ISL_413609 | Nasopharyngeal swab  | 2020-02-21     | C1385T, C29230T, C29635T                                                 |
| hCoV-19/USA/CruiseA-5/2020 | 5    | EPI_ISL_413610 | Oropharyngeal swab   | 2020-02-21     |                                                                           |
| hCoV-19/USA/CruiseA-6/2020 | 6    | EPI_ISL_413611 | Nasopharyngeal swab  | 2020-02-21     | G11410A, C26326T                                                         |
| hCoV-19/USA/CruiseA-7/2020 | 7    | EPI_ISL_413612 | Nasopharyngeal swab  | 2020-02-17     | C3738T (Orf1a P1158L), G11410A, C26326T                                   |
| hCoV-19/USA/CruiseA-8/2020 | 8    | EPI_ISL_413613 | Nasopharyngeal swab  | 2020-02-17     | C9474T (Orf1ab A3070V)                                                   |
| hCoV-19/USA/CruiseA-9/2020 | 9    | EPI_ISL_413614 | Nasopharyngeal swab  | 2020-02-17     |                                                                           |
| hCoV-19/USA/CruiseA-10/2020| 10   | EPI_ISL_413615 | Nasopharyngeal swab  | 2020-02-17     | G3259T (Orf1ab Q998H)                                                    |
| hCoV-19/USA/CruiseA-11/2020| 11   | EPI_ISL_413616 | Nasopharyngeal swab  | 2020-02-17     | C10036T                                                                  |
| hCoV-19/USA/CruiseA-12/2020| 12   | EPI_ISL_413617 | Oropharyngeal swab   | 2020-02-20     | C29635T                                                                  |
| hCoV-19/USA/CruiseA-13/2020| 13   | EPI_ISL_413618 | Nasopharyngeal swab  | 2020-02-20     |                                                                           |
| hCoV-19/USA/CruiseA-14/2020| 14   | EPI_ISL_413619 | Oropharyngeal swab   | 2020-02-25     | C6636T (Orf1ab T2124I), C11750T, C11956T                                  |
| hCoV-19/USA/CruiseA-15/2020| 15   | EPI_ISL_413620 | Nasopharyngeal swab  | 2020-02-18     |                                                                           |
| hCoV-19/USA/CruiseA-16/2020| 16   | EPI_ISL_413621 | Nasopharyngeal swab  | 2020-02-18     |                                                                           |
| Name | EPI ISL 413622 | Type | Date | Sequence Details |
|------|----------------|------|------|----------------|
| hCoV-19/USA/CruiseA-17/2020 | EPI_ISL_413623 | Nasopharyngeal swab | 2020-02-24 | Deletion 508-22, C22033A (S protein F156L) |
| hCoV-19/USA/CruiseA-19/2020 | EPI_ISL_414479 | Nasopharyngeal swab | 2020-02-18 | |
| hCoV-19/USA/CruiseA-21/2020 | EPI_ISL_414480 | Oropharyngeal swab | 2020-02-21 | C254T, G29726T, G29751C |
| hCoV-19/USA/CruiseA-22/2020 | EPI_ISL_414481 | Nasopharyngeal swab | 2020-02-21 | |
| hCoV-19/USA/CruiseA-23/2020 | EPI_ISL_414482 | Nasopharyngeal swab | 2020-02-18 | C254T, T9157C, G11083T (Orf1ab L3606F), G22104T (S protein G181V), G29736T, G29751C |
| hCoV-19/USA/CruiseA-24/2020 | EPI_ISL_414483 | Oropharyngeal swab | 2020-02-17 | C3099T (Orf1a T945I), C10507T, G11083T (Orf1ab L3606F), G28378T |
| hCoV-19/USA/CruiseA-25/2020 | EPI_ISL_414484 | Nasopharyngeal swab | 2020-02-17 | G11083T (Orf1ab L3606F) |
| hCoV-19/USA/CruiseA-26/2020 | EPI-ISL_414485 | Oropharyngeal swab | 2020-02-24 | G11083T (Orf1ab L3606F), G28916A (N protein G214S) |
| hCoV-19/Japan/TK/20-31-3/2020 | EPI_ISL_413459 | Autopsy (deceased), Bronch and lung | 2020-02-20 | G11083T (Orf1ab L3606F), C29635T |
| hCoV-19/Japan/Hu_D_P_Kng_19-027/2020 | EPI-ISL_412969 | Throat swab | 2020-02-10 | G11083T (Orf1ab L3606F), C29635T |
| hCoV-19/Japan/Hu_D_P_Kng_19-020/2020 | EPI-ISL_412968 | Throat swab | 2020-02-10 | G11083T (Orf1ab L3606F) |