

Materials and Methods

Clinical specimens and RT-qPCR testing for COVID-19 infection

To permit the passengers to disembark, a quantitative reverse transcription PCR (RT-qPCR) test for SARS-CoV-2 (*12, 13*) was performed. In this study, we obtained the pharyngeal specimens of 896 persons (880 passengers, 15 crew members, and 1 quarantine officer) from the Diamond Princess (DP) during Feb. 15–17 before disembarking on Feb. 19, 2020. This study was approved by the research ethics committee of the National Institute of Infectious Diseases (approval no. 1091).

Whole genome sequencing of SARS-CoV-2

To elucidate the route of dissemination on the DP during the quarantine period, we conducted whole genome sequencing (WGS) of SARS-CoV-2 using a multiplex PCR-based RNA-Seq (PrimalSeq) (*14, 15*) technique directly from PCR-positive clinical specimens. WGS analysis provides useful information for tracing back infection linkages. However, the virus content in pharyngeal swabs is not abundant at the early onset of COVID-19, which might lead to difficulties in achieving WGS. Indeed, the genome size of SARS-CoV-2 is 29.9 kb, which is the largest among known RNA viruses, suggesting that some improvements will be required to enrich the pathogen RNA molecule at every preparation step of the sequencing protocol to determine the WGS efficiently.

The whole genome sequences of SARS-CoV-2 isolates were obtained by the PrimalSeq protocol for producing enriched cDNA of the SARS-CoV-2 genome by multiplex RT-PCR amplicons using a multiplexed PCR primer set which was proposed by Wellcome Trust ARTIC Network (*16*). The PCR products from the same clinical samples were pooled, purified, and

subjected to Illumina library construction using QIAseq FX DNA Library Kit (QIAGEN, Hilden Germany). The NextSeq 500 platform (Illumina, San Diego, USA) was used for sequencing the indexed libraries. We found that two particular amplicons regularly showed low to zero coverage due to primer dimerization as described in Itokawa et al (17) (we should note that the modified primer was not used in this study). Then, we used data from another round of 3 kb RT-PCR assays covering the whole genome of SARS-CoV-2 (29.8 kb) with ten PCR amplicons. The 3 kb RT-PCR multiplexed amplicons were mixed in a single tube and subjected to shotgun next-generation sequencing (NGS) on the NextSeq 500 platform (Illumina, San Diego, USA) as described for the ARTIC library. The NGS reads were mapped to the SARS-CoV-2 Wuhan-Hu-1 reference genome sequence (29.9 kb ss-RNA; GenBank ID: MN908947), resulting in a specimen-specific SARS-CoV-2 genome sequence fully mapped to the reference. The sample sequences were assembled using A5-miseq v.20140604 (18).

Comparative genome sequence analysis and single nucleotide variation analysis

The nearly full-length genome sequence (≥ 29 kb) of SARS-CoV-2 were retrieved from GISAID EpiCoV database in March 10, 2020, followed by multiple alignment using MAFFT v7.222 (19). The poorly aligned regions in 5' and 3' end were trimmed; we determined that the core regions were from 99 to 29794 nt position against Wuhan-Hu-1 genome sequences (GISAID ID, EPI_ISL_402125; GenBank ID, MN908947.3). Gap-containing sequences in the core region were excluded; sequences of 243 isolates in GISAID database were eventually used in subsequent analyses. The genome sequences of DP isolates were aligned using MAFFT program together with sequences retrieved from database, followed by extraction of SNV and deletion sites. The SNV median-joining network analysis was performed by PopART software (20).

Table S1.

RT-qPCR and whole genome sequence information in this study.
See Supplementary_Table_S1.xlsx

Table S2.

Genome-wide single nucleotide variations of SARS-CoV-2 isolates in this study
See Supplementary_Table_S2.xlsx

Table S3.

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database on which this research is based.
See Supplementary_Table_S3.xlsx