Infection Prevention in Practice 3 (2021) 100190

Available online at www.sciencedirect.com

Infection Prevention in Practice

journal homepage: www.elsevier.com/locate/ipip

# Genetic epidemiology using whole genome sequencing and haplotype networks revealed the linkage of SARS-CoV-2 infection in nosocomial outbreak

**Fumihiro Ishikawa** [a,b,*], **Yuko Udaka** [a,c], **Hideto Oyamada** [a,c], **Keiko Ishino** [a,d], **Issei Tokimatsu** [e], **Hironori Sagara** [f], **Yuji Kiuchi** [a,c]

[a] *PCR Centre for COVID-19, Showa University Hospital, Tokyo, Japan*
[b] *Centre for Biotechnology, Showa University, Tokyo, Japan*
[c] *Department of Pharmacology, Showa University School of Medicine, Tokyo, Japan*
[d] *Division of Infection Control Sciences, Department of Clinical Pharmacy, Showa University School of Pharmacy, Tokyo, Japan*
[e] *Division of Infectious Diseases, Department of Internal Medicine, Showa University School of Medicine, Tokyo, Japan*
[f] *Division of Respiratory Medicine and Allergology, Department of Internal Medicine, Showa University School of Medicine, Tokyo, Japan*

ARTICLE INFO

*Article history:*
Received 3 August 2021
Accepted 17 November 2021
Available online 24 November 2021

*Keywords:*
Epidemiological analysis
Haplotype networks
Phylogenetic tree analysis
Severe acute respiratory syndrome coronavirus 2
Whole genome sequencing

SUMMARY

***Background:*** A characteristic feature of SARS-CoV-2 is its ability to transmit from pre- or asymptomatic patients, complicating the tracing of infection pathways and causing outbreaks. Despite several reports that whole genome sequencing (WGS) and haplotype networks are useful for epidemiologic analysis, little is known about their use in nosocomial infections.

***Aim:*** We aimed to demonstrate the advantages of genetic epidemiology in identifying the link in nosocomial infection by comparing single nucleotide variations (SNVs) of isolates from patients associated with an outbreak in Showa University Hospital.

***Methods:*** We used specimens from 32 patients in whom COVID-19 had been diagnosed using clinical reverse transcription-polymerase chain reaction tests. RNA of SARS-CoV-2 from specimens was reverse-transcribed and analysed using WGS. SNVs were extracted and used for lineage determination, phylogenetic tree analysis, and median-joining analysis.

***Findings:*** The lineage of SARS-CoV-2 that was associated with outbreak in Showa University Hospital was B.1.1.214, which was consistent with that found in the Kanto metropolitan area during the same period. Consistent with canonical epidemiological observations, haplotype network analysis was successful for the classification of patients. Additionally, phylogenetic tree analysis revealed three independent introductions of the virus into the hospital during the outbreak. Further, median-joining analysis indicated that four patients were directly infected by any of the others in the same cluster.

*Abbreviations:* COVID-19, coronavirus disease 2019; WGS, whole genome sequencing; SNVs, single nucleotide variations; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SUH, Showa University Hospital; RT-PCR, reverse transcription-polymerase chain reaction; GISAID, Global Initiative on Sharing All Influenza Data; KMA, Kanto metropolitan area; NJ, neighbour-joining; VOC, variant of concern.

* Corresponding author. Address: Centre for Biotechnology, Showa University, 1-5-8 Hatanodai, Shinagawa-ku, Tokyo, 142-8555, Japan. Tel.: +81 3 3784-8243 (mobile).
*E-mail address:* f-ishikawa@pharm.showa-u.ac.jp (F. Ishikawa).

https://doi.org/10.1016/j.infpip.2021.100190

***Conclusion:*** Genetic epidemiology with WGS and haplotype networks is useful for tracing transmission and optimizing prevention strategies in nosocomial outbreaks.



## Introduction

Coronavirus disease 2019 (COVID-19) originated in Wuhan, China, in late December 2019, rapidly spread worldwide, and was officially declared a pandemic by the World Health Organization on 11th March 2020 [1]. COVID-19 patients can infect other individuals before symptom onset or even without development of any apparent symptoms [2,3]. This pre- and asymptomatic disease transmission makes it difficult to identify the origin of infection and prevent further spread, leading to disease outbreaks.

COVID-19 is caused by SARS-CoV-2, a single-stranded RNA virus that rapidly accumulates mutations in its genome during replication [4]. Taking advantage of the high frequency of mutations, haplotypes with single nucleotide variations (SNVs) have been utilized for epidemiological analysis, supported by whole genome sequencing (WGS). To date, many studies have reported the advantage of haplotype network analysis with WGS-SNVs in understanding the evolution of SARS-CoV-2 and the origin of its international spread [5–7]. Similarly, in the case of nosocomial outbreaks, Takenouchi *et al*. reported that WGS was useful in determining whether sporadic cases were indeed part of an institutional cluster [8]. However, there are few studies supporting the usefulness of WGS-SNVs in nosocomial outbreaks for infection control.

Showa University Hospital (SUH) experienced a COVID-19 outbreak between mid-December 2020 and late January 2021. In this study, we aimed to analyse the phylogenetic lineage of the virus population detected in SUH and to evaluate the transmission between COVID-19 patients in this outbreak using haplotype network analysis.

## Methods

### *Study population*

This study was conducted using specimens from 32 patients in whom COVID-19 had been diagnosed by clinical reverse transcription-polymerase chain reaction (RT-PCR) test at SUH between 1st July 2020 and 31st January 2021. These patients included 11 healthcare workers and 21 patients, designated SUH001 to SUH032. This study did not include personal information leading to identification of individuals.

The study protocol was approved by the ethics committee of Showa University School of Medicine (approval number: 3302).

### *Specimens and clinical RT-PCR test for SARS-CoV-2 infection*

Nasopharyngeal swabs were collected from suspected COVID-19 patients and tested using quantitative reverse transcription polymerase chain reaction (RT-qPCR) performed using a SARS-CoV-2 detection kit (TOYOBO, Osaka, Japan), according to the manufacturer's instructions.

### *RNA extraction and whole genome sequencing*

Viral RNA was extracted from residual specimens after clinical RT-qPCR using the QIAamp Viral RNA Mini Kit (QIAGEN, Hilden, Germany). A sequencing library for the whole genome of SARS-CoV-2 isolates was prepared according to the nCoV-2019 sequencing protocol for Illumina V.2 (https://www.protocols.io/view/ncov-2019-sequencing-protocol-for-illumina-betejeje?version_warning=no). Quantification of the amplicon pools was performed using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA), and paired-end sequencing was performed on the MiSeq next-generation sequencing (NGS) platform (Illumina, San Diego, CA, USA). The reads obtained from NGS were mapped to a reference sequence (Wuhan-Hu-1, GenBank ID: NM908947.3) using Minimap2 version 2.17 [9], and the resulting alignment sequence was trimmed at both ends to remove multiplex primer sequences using iVar version 1 [10]. All the WGS data obtained have been deposited in the Global Initiative on Sharing All Influenza Data (GISAID) database (https://www.gisaid.org/). The GISAID IDs are presented in Appendix A.

### *Bioinformatic analysis*

We retrieved the whole genome sequences of SARS-CoV-2 detected in the Kanto metropolitan area (KMA) between 1st July 2020 and 31st January 2021 from the GISAID EpiCoV database (N = 577). Multiple alignment was conducted by comparing a reference sequence (Wuhan-Hu-1, GenBank ID: NM908947.3) with sequences retrieved from the GISAID database, using MAFFT version 7 software [11]. The regions corresponding to 33–29866 nt of Wuhan-Hu-1 were defined as the core region and used for further analysis. The mixed bases in the core region were resolved by manually counting the sequence reads. A phylogenetic tree analysis with SNVs was performed using the neighbour-joining (NJ) method and MAFFT version 7 [11], followed by visualization with iTOL version 6 [12]. The median-joining network analysis with SNV was performed using PopART software [13]. PANGO lineages of isolates were examined by phylogenetic assignment of named global outbreak lineages web application (Pangolin) (https://pangolin.cog-uk.io/) [14]. A bubble chart of the PANGO lineage was created using JMP Pro version 15 (SAS Institute Inc., Cary, NC, USA).

## Results

### *Phylogenetic analysis of SARS-CoV-2 isolates in KMA and SUH*

SARS-CoV-2 rapidly accumulates mutations in the genome, giving rise to a new lineage. As of 25th June 2021, more than 1500 lineages have been reported on the PANGO lineageswebsite (https://cov-lineages.org/). Since viral transmissibility is

dependent on their lineages, we compared the phylogenetic classification of SARS-CoV-2 isolates obtained at SUH to that of isolates detected in KMA during the same time period (Figure 1). The KMA and SUH isolates were classified into nine lineages, all of which have *A23403G* substitution in the genomes, resulting in the D614G mutation in the spike protein (Table I). This observation indicated that all isolates were progenies of the virus related to the European outbreak in early 2020 but had not first been introduced from China [7]. More than 90% of the isolates belonged to B.1.1.284 and B.1.1.214 lineages unique to Japan, which were responsible for the outbreak in 2020. There was no difference in the lineages between the SUH and KMA isolates, suggesting that the outbreak in SUH was not caused by isolates with higher transmissibility compared to those in other hospitals.

### *Haplotype network analysis with SNVs in SUH hospital outbreak*

To clarify the linkage of SARS-CoV-2 infection in the SUH outbreak, we used haplotype network analysis. The SUH nosocomial outbreak consisted of nearly 40 patients in four wards between mid-December 2020 and late January 2021, and 17 of outbreak isolates (SUH013 to SUH022, SUH024 to SUH026, SUH028, and SUH030 to SUH032) were included in this study. To assess the haplotype network analysis, we included the following isolates predicted to be irrelevant to the nosocomial outbreak that we had experienced between mid-December 2020 and late January 2021 in the analysis: isolates detected during different periods from July to November (SUH001 to SUH009) and those from community-acquired outpatients with COVID-19 or patients transferred from another hospital (SUH010 to SUH012, SUH023, SUH027, and SUH029) during the same period. Using these isolates as a control, we performed multiple alignment analysis using Wuhan-Hu-1 as a reference sequence and identified 107 SNVs in genomes of 32 SUH isolates, and the average was 17.3 (standard deviation, 2.0) per isolate. In addition, more than half of the SNVs (64) were specific for each isolate, indicating that there could be sufficient diversity for discriminating between the isolates.

To investigate the evolutionary connection among isolates, we performed a phylogenetic tree analysis, which found four clades that contained multiple isolates (Figure 2). Clade 1 was composed of five isolates, all of which belonged to lineage B.1.1.284. Patients SUH011, SUH012, and SUH023 were expected to have close contact with each other because they had lived in the same house/room. Correspondingly, all isolates from these three patients resided in clade 2 and were characterized by five SNVs (*C1684T*, *C9207T*, *C15240T*, *T22447C*, and *T24469C*). In contrast, the isolates used as controls were in different clades. These observations indicate that this phylogenetic analysis successfully classified isolates according to their genomic sequence diversity. Under these conditions, isolates related to nosocomial outbreaks were separated into clades 3 and 4 without SUH028. Clades 3 and 4 had unique SARS-CoV-2 haplotypes characterized by three SNVs (*C9803T*, *C16887T*, and *G25947T*) and four SNVs (*A594G*, *G11804A*, *C27630T*, and *C27881T*), respectively. The presence of clade 3 and clade 4 indicates two independent subclusters (named cluster 1 and cluster 2, respectively). SUH028 was distinguished from the other isolates by its nine SNVs (*A385G*, *C5284T*, *C6380T*, *C7732T*, *C15981T*, *C17745T*, *C23757T*, *C26607T*, *C29409T*), which could be potentially shared with isolates that were not included in this study. These data suggest that there were at least three introductions of SARS-CoV-2 into SUH during the period of the outbreak.

In nosocomial infections, identification of the infection link in a cluster is extremely important to design preventive strategies against subsequent spread. Although phylogenetic tree analysis is useful for the identification of groups as clades of isolates that have similar genomic sequences, it is difficult to

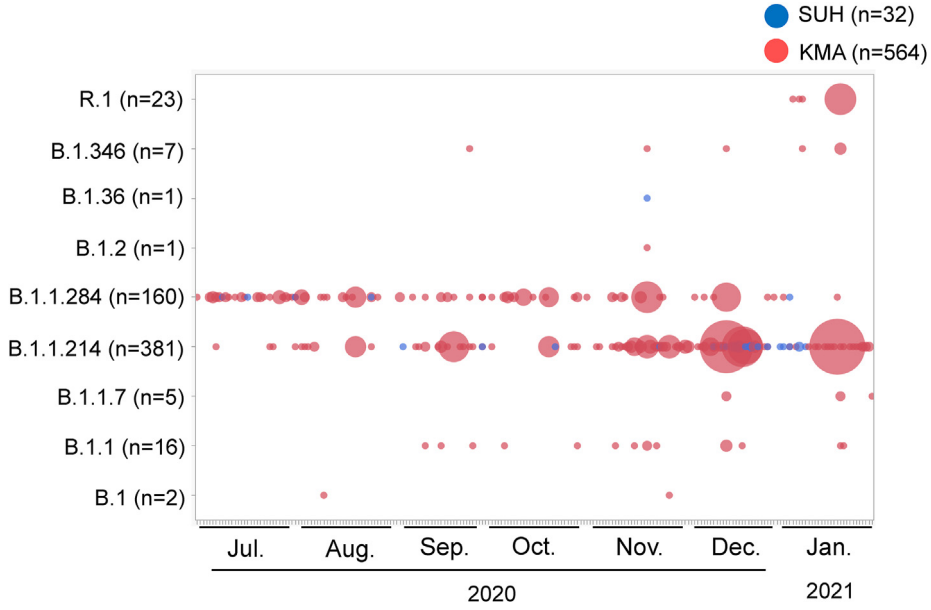


**Figure 1.** Phylogenetic classification of SARS-CoV-2 genome from KMA. Nine lineages were detected in 596 isolates from KMA, including the SUH (n = 32) samples, during the evaluation period, as visualized with a bubble plot. The red and blue bubbles represent the isolates from KMA and SUH, respectively. The size of the bubbles is proportional to the number of detected isolates.

**Table I**
SNVs of 32 isolates detected in SUH

| Patients | SNVs detected |
|---|---|
| SUH001 | C313T, G2167T, C3037T, T4346C, C9286T, C10376T, C14408T, C14708T, T22020C, A23403G, C28725T, G28881A, G28882A, G28883C, G29692T |
| SUH002 | C313T, C3037T, T4346C, C9286T, A10269C, C10376T, C14408T, C14708T, T22020C, A23403G, C28725T, G28881A, G28882A, G28883C, G29692T |
| SUH003 | C313T, C1108T, C3037T, T4346C, C9286T, C10376T, C14408T, C14708T, T22020C, A23403G, C28725T, G28881A, G28882A, G28883C, C29077T, G29692T |
| SUH004 | C313T, C3037T, T4346C, C9286T, C10376T, C14408T, C14708T, T22020C, A23403G, C23988T, C28725T, G28881A, G28882A, G28883C, G29692T |
| SUH005 | C313T, A2368G, C3037T, C6433T, C14408T, T15597C, C18167T, G21518T, A23403G, G28881A, G28882A, G28883C, G28975T |
| SUH006 | C313T, C3037T, C4331T, C8917T, G11335T, C14408T, C16650T, C18167T, G21518T, C23127T, A23403G, G28881A, G28882A, G28883C, G28975T, G29560T, C29679T, C29831T |
| SUH007 | C313T, C934T, C3037T, C6433T, C14408T, G15921T, G16158T, G17695T, C18167T, G21518T, A23403G, C28333T, G28881A, G28882A, G28883C, G28975T |
| SUH008 | C1326T, T1947C, C3037T, C5467T, G9190T, C9891T, C14408T, C18877T, C21855T, C22444T, C22978T, A23403G, G25563T, C25728T, C26735T, C28854T |
| SUH009 | C241T, C313T, A1643G, A2861G, C3037T, G8371T, C8917T, C14408T, C18167T, A18550G, G19656A, G21518T, A23403G, G26428T, C27509T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH010 | C313T, C3037T, C6433T, G8371T, C11776T, C13019T, C14408T, C15240T, C18167T, G21305A, G21518T, C22564T, A23403G, G28881A, G28882A, G28883C, G28975T, C29523T |
| SUH011 | C313T, C1684T, C3037T, C6433T, G8371T, C9207T, C14408T, C15240T, C18167T, G21518T, T22447C, A23403G, T24469C, G28881A, G28882A, G28883C, G28975T |
| SUH012 | C313T, C1684T, C3037T, C6433T, C9207T, C14408T, C15240T, C18167T, G21518T, T22447C, A23403G, T24469C, G28881A, G28882A, G28883C, G28975T |
| SUH013 | C313T, A594G, C3037T, C7321T, C8917T, C10156T, G11804A, C14408T, C18167T, G21518T, A23403G, C27630T, C27881T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH014 | C313T, A594G, C3037T, C7321T, C8917T, C10156T, G11804A, C14408T, C18167T, G21518T, A23403G, C27630T, C27881T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH015 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH016 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH017 | C313T, A594G, C3037T, C7321T, C8917T, C10156T, G11804A, C14408T, C18167T, G21518T, A23403G, C27630T, C27881T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH018 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH019 | C313T, A594G, C3037T, C7321T, C8917T, C10156T, G11804A, C14408T, C18167T, G21518T, A23403G, C27630T, C27881T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH020 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH021 | C313T, A594G, C3037T, C7321T, C8917T, C10156T, G11804A, C14408T, C18167T, G21518T, A23403G, C27630T, C27881T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH022 | C313T, C1044T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH023 | C313T, C1684T, C3037T, C6433T, G8371T, C9207T, C14408T, C15240T, C18167T, G21518T, T22447C, A23403G, T24469C, G28881A, G28882A, G28883C, G28975T |
| SUH024 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18060T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH025 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH026 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, C18312T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH027 | C313T, C3037T, G8602T, C8917T, C14408T, C18167T, G19072T, C20844T, G21518T, A23403G, G28881A, G28882A, G28883C, G28975T |
| SUH028 | C313T, A385G, C3037T, C5284T, C6380T, C7732T, C8917T, C14408T, C15981T, C17745T, C18167T, A18550G, G19656A, G21518T, A23403G, C23757T, C26607T, G28881A, G28882A, G28883C, G28975T, C29409T, C29679T |

**Table I** (*continued*)

| Patients | SNVs detected |
|---|---|
| SUH029 | C313T, C3037T, T4346C, C9286T, A9343G, C10376T, T12145G, C14408T, C14708T, A17008G, T22020C, A23403G, G25996T, A27633G, G28703C, C28725T, G28881A, G28882A, G28883C, G29692T |
| SUH030 | C313T, C3037T, C7321T, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH031 | C313T, C3037T, C7321T, C7573A, C8917T, C9803T, C10156T, C14408T, C16887T, C18167T, G21518T, A23403G, G25947T, G28881A, G28882A, G28883C, G28975T, C29679T |
| SUH032 | C313T, A594G, C3037T, C7321T, C8917T, C10156T, G10318T, G11804A, C14408T, C18167T, C18646T, G21518T, A23403G, C25469T, C27630T, C27881T, G28881A, G28882A, G28883C, G28975T, C29679T |

reveal the parent-child relationship of isolates. Median-joining network analysis has shown excellent results in this regard because it can graphically express the difference in a single nucleotide in the genome [6,15]. Therefore, to gain a deeper insight, we performed a median-joining network analysis using SNVs of the SUH isolates (Figure 3). In cluster 1 (n = 10), six isolates had identical genome sequences (Figure 3, vertex A) and other isolates had one unique SNV individually. These data indicate that four patients had COVID-19 by direct infection from any of the other patients in vertex A. Meanwhile, the genome sequences of five isolates were identical (Figure 3, vertex B), and one isolate showed an additional three SNVs in cluster 2 (n = 6) as phylogenetic tree analysis. As expected, all isolates from patients with a close contact history (n = 3) were identical (Figure 3, vertex C).

## Discussion

In this study, we evaluated the advantages of using WGS and haplotype network analysis to identify the infection link in nosocomial outbreaks. First, we examined the SARS-CoV-2 lineages that were related to the SUH outbreak and determined it to be B.1.1.214. To date, multiple variants of concern (VOCs) that have a higher effective reproduction number and an increased risk of mortality have been spreading worldwide as pandemic strains [16,17]. We did not find any VOCs, such as the alpha variant B.1.1.7. Currently, the delta variant B.1.617.2 and its subvariants have been expanding in many countries, threatening healthcare systems by rapidly increasing the number of symptomatic patients. Because the introduction of VOCs into a hospital could easily cause an outbreak, the identification of COVID-19-positive patients and their viral strains are necessary for quick and decisive actions, such as single-patient room management, to save lives. The WGS by NGS is essential to determine the lineage, although each mutation of the SARS-CoV-2 genome can be detected by single nucleotide polymorphism genotyping [18].

Our results showed that phylogenetic tree analysis could clearly distinguish between patients infected in the hospital and those infected in the community as well as classify patients with a close contact history as a single clade. This agrees with findings of a previous study [8], which also indicated that phylogenetic tree analysis was successful in discriminating between COVID-19 patients on the basis of the genomic sequence of isolates during the outbreak. Surprisingly, our data showed that three separate introductions of SARS-CoV-2 during the same period into the hospital had played individual parts in the outbreak. As only patients who had tested negative for SARS-CoV-2 by RT-qPCR several days before admission were in the hospital, these introductions may have been due to healthcare workers or patients infected with a viral load below the detection limit. Additionally, one patient in cluster 1 was hospitalized on a different floor and had no direct contact with other patients in cluster 1, suggesting transmission either through healthcare workers or in shared spaces, such as an elevator. This result raises the possibility that even the adherence to standard precautions in healthcare workers when providing patient care and disinfection of shared space was insufficient to prevent transmission. Therefore, healthcare workers must pay more attention to adequate precautions.

Notably, we showed that median-joining network analysis with SNVs could indicate the direction of transmission. It should be noted that while we could identify the individuals involved in the dissemination of this virus using this method, it is difficult to estimate the direction when the SNVs are identical. To resolve this issue, applying canonical epidemiological observations such as the date of symptom onset may be helpful. Identification of individuals responsible for the transmission would clarify how the transmission occurred by intensively investigating their recent activities, leading to improvement of infection control. Meanwhile, as Johnson and Parker pointed out, information on transmission during the outbreak could have potentially harmful consequences when the source of the outbreak is identified [19]. These consequences could include psychological distress and affording the responsibility of transmission to individuals. Therefore, the sharing and use of data should also be based on ethical considerations.

The main limitations of this study include its retrospective nature and the lack of evaluation of haplotype network analysis using WGS-SNVs for infection control. Rapid onsite WGS during an outbreak and adequate intervention to prevent further spread are necessary for evaluation of this strategy. However, we could not accomplish this due to our lack of NGS equipment and consignment of WGS operation to an external institution. Because it takes several months at the earliest to receive sequences from the external institution, intervention during an outbreak would be difficult. Whether the interventions implemented to block the infection pathways identified by haplotype networks is successful for infection control should be

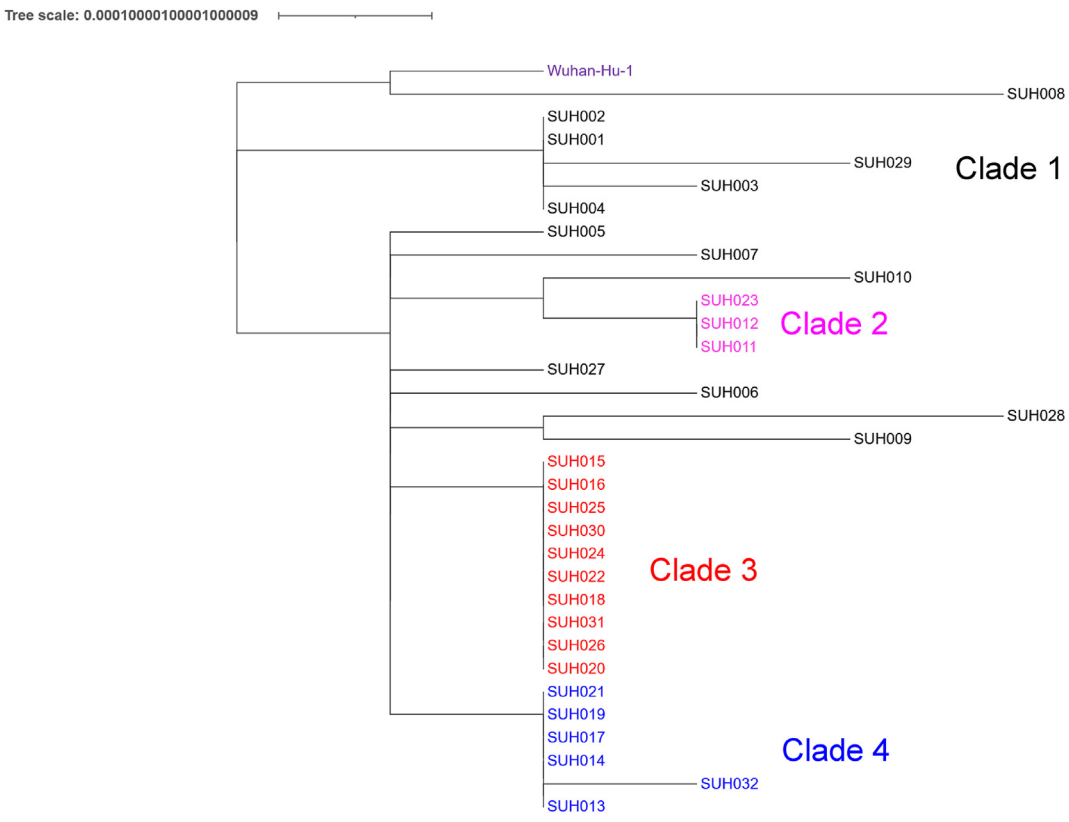


**Figure 2.** Phylogenetic tree analysis of SARS-CoV-2 genome from SUH. A phylogenetic tree was created using the genomes of the 32 isolates from SUH with the NJ method, as described in the Methods section.

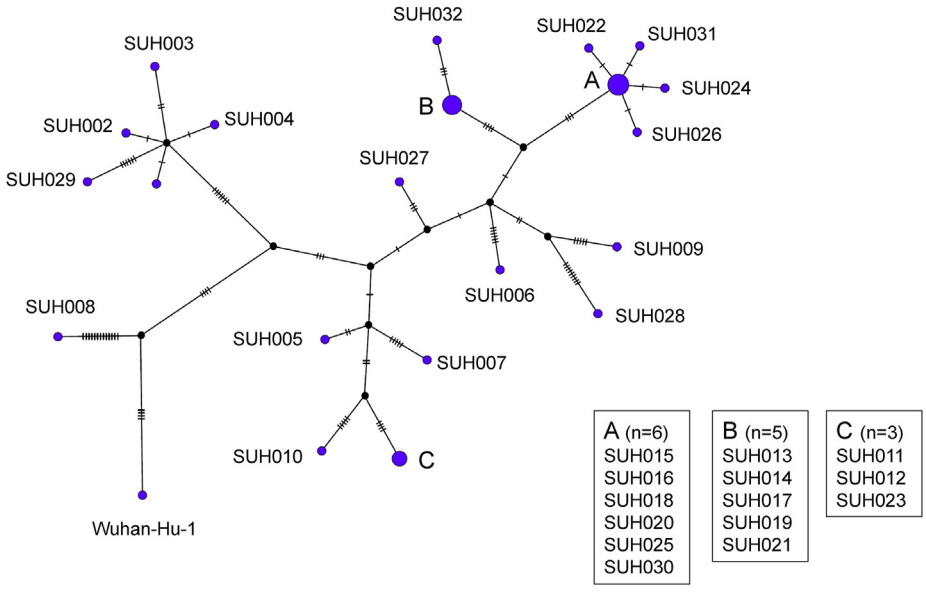


**Figure 3.** Median-joining network analysis of SARS-CoV-2 genome from SUH. The median-joining network analysis with SNV was performed using the isolates as in Figure 2. The isolates from SUH patients are shown as blue circles, and their sizes are proportional to the number of isolates. The number of hatch marks indicates the number of SNVs between the isolates.

addressed by future studies performed in hospitals that have NGS equipment. In the past two decades, three types of coronavirus have emerged and caused outbreaks in many countries [20]. This indicates the possibility that another coronavirus associated with outbreak could emerge in the near future. In such an event, it might be important to perform in-hospital WGS to promptly end a nosocomial outbreak.

## Conclusions

Identification of infection links is a crucial step in infection control because it assists in determining pathogen transmission and in designing strategies to prevent further spread of infection. Our results provide evidence that genetic epidemiology with WGS and haplotype networks are useful for assisting canonical epidemiology by tracing transmission and optimizing prevention strategies in nosocomial outbreaks.

## CRediT author statement

**Fumihiro Ishikawa** : Conceptualization, Investigation, Writing - Original Draft.
**Yuko Udaka:** Investigation.
**Hideto Oyamada:** Investigation.
**Keiko Ishino:** Investigation.
**Issei Tokimatsu:** Supervision.
**Hironori Sagara:** Supervision.
**Yuji Kiuchi:** Project administration.

## Acknowledgements

We thank all the patients and medical staff who have participated in this study, the supporting staff of the PCR Centre for their contribution to the clinical RT-PCR testing, TAKARA Bio (Shiga, Japan) for NGS operation, Editage (www.editage.com) for English language editing, and all authors who have kindly deposited genome data used in this study on the GISAID database.

## Conflict of interest statement

The authors do not have any conflicts of interest to declare.

## Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Appendices

### *Appendix A IDs deposited in GISAID*

EPI_ISL_1716880, EPI_ISL_1752589 to EPI_ISL_1752613, EPI_ISL_1752615, E1752617, EPI_ISL_1752633 to EPI_ISL_1752635.

### *Appendix B IDs retrieved from GISAID*

EPI_ISL_644955 to EPI_ISL_644999, EPI_ISL_693298, EPI_ISL_693299, EPI_ISL_803879, EPI_ISL_860117 to EPI_ISL_860181, EPI_ISL_915358 to EPI_ISL_915384, EPI_ISL_915389 to EPI_ISL_915420, EPI_ISL_962522, EPI_ISL_1034196 to EPI_ISL_1034208, EPI_ISL_1034212 to EPI_ISL_1034226, EPI_ISL_1041931 to EPI_ISL_1041955, EPI_ISL_1069157 to EPI_ISL_1069171, EPI_ISL_1072966 to EPI_ISL_1072984, EPI_ISL_1078589, EPI_ISL_1078590, EPI_ISL_1078592 to EPI_ISL_1078594, EPI_ISL_1078596, EPI_ISL_1078597, EPI_ISL_1078599 to EPI_ISL_1078601, EPI_ISL_1078603, EPI_ISL_1078604, EPI_ISL_1078606 to EPI_ISL_1078608, EPI_ISL_1078610,EPI_ISL_1078611, EPI_ISL_1078613, EPI_ISL_1078614, EPI_ISL_1078616 to EPI_ISL_1078618, EPI_ISL_1078620, EPI_ISL_1078621, EPI_ISL_1123300 to EPI_ISL_1123324, EPI_ISL_1123331, EPI_ISL_1123412, EPI_ISL_1125391, EPI_ISL_1125393, EPI_ISL_1125395, EPI_ISL_1127093 to EPI_ISL_1127105, EPI_ISL_1127112 to EPI_ISL_1127117, EPI_ISL_1129228 to EPI_ISL_1129242, EPI_ISL_1137610, EPI_ISL_1172040, EPI_ISL_1172041, EPI_ISL_1172045, EPI_ISL_1448044 to EPI_ISL_1448047, EPI_ISL_1761308 to EPI_ISL_1761350, EPI_ISL_1761363, EPI_ISL_1973091 to EPI_ISL_1973127, EPI_ISL_1973131 to EPI_ISL_1973160, EPI_ISL_1973162 to EPI_ISL_1973171, EPI_ISL_1973186 to EPI_ISL_1973202, EPI_ISL_1973204 to EPI_ISL_1973233, EPI_ISL_1973240 to EPI_ISL_1973248, EPI_ISL_1973250 to EPI_ISL_1973253, EPI_ISL_1973255, EPI_ISL_1973267, EPI_ISL_2285724, EPI_ISL_2285725, EPI_ISL_2285730, EPI_ISL_2285731, EPI_ISL_2303990 to EPI_ISL_2303997, EPI_ISL_2321212 to EPI_ISL_2321225.

## References

[1] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265—9. https://doi.org/10.1038/s41586-020-2008-3.
[2] Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA 2020;323:1406—7. https://doi.org/10.1001/jama.2020.2565.
[3] Byambasuren O, Cardona M, Bell K, Clark J, McLaws M-L, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. Off J Assoc Med Microbiol Infect Dis Can 2020;5:223—34. https://doi.org/10.3138/jammi-2020-0030.
[4] Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell 2020;181:914—921.e10. https://doi.org/10.1016/j.cell.2020.04.011.
[5] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl. Acad Sci U S A 2020;117:9241—3. https://doi.org/10.1073/pnas.2004999117.
[6] Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, et al. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. Proc Natl. Acad Sci U S A 2020;117:20198—201. https://doi.org/10.1073/pnas.2006824117.
[7] Sekizuka T, Itokawa K, Hashino M, Kawano-Sugaya T, Tanaka R, Yatsu K, et al. A genome epidemiological study of SARS-CoV-2 introduction into Japan. mSphere 2020;5. https://doi.org/10.1128/mSphere.00786-20.
[8] Takenouchi T, Iwasaki YW, Harada S, Ishizu H, Uwamino Y, Uno S, et al. Clinical utility of SARS-CoV-2 whole genome sequencing in deciphering source of infection. J Hosp Infect 2021;107:40—4. https://doi.org/10.1016/j.jhin.2020.10.014.
[9] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094—100. https://doi.org/10.1093/bioinformatics/bty191.
[10] Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol 2019;20:8. https://doi.org/10.1186/s13059-018-1618-7.

[11] Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 2019;20:1160—6. https://doi.org/10.1093/bib/bbx108.

[12] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 2019;47:W256—9. https://doi.org/10.1093/nar/gkz239.

[13] Leigh JW, Bryant D, Nakagawa S. Popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution 2015;6:1110—6. https://doi.org/10.1111/2041-210x.12410.

[14] Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403—7. https://doi.org/10.1038/s41564-020-0770-5.

[15] Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999;16:37—48. https://doi.org/10.1093/oxfordjournals.molbev.a026036.

[16] Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. BMJ 2021;372:n579. https://doi.org/10.1136/bmj.n579.

[17] Sheikh A, McMenamin J, Taylor B, Robertson C. Public Health Scotland and the EAVE II Collaborators. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. Lancet 2021;397:2461—2. https://doi.org/10.1016/S0140-6736(21)01358-1.

[18] Harper H, Burridge A, Winfield M, Finn A, Davidson A, Matthews D, et al. Detecting SARS-CoV-2 variants with SNP genotyping. PLOS ONE 2021;16:e0243185. https://doi.org/10.1371/journal.pone.0243185.

[19] Johnson SB, Parker M. The ethics of sequencing infectious disease pathogens for clinical and public health. Nat Rev Genet 2019;20:313—5. https://doi.org/10.1038/s41576-019-0109-3.

[20] Guarner J. Three emerging coronaviruses in two decades. Am J Clin Pathol 2020;153:420—1. https://doi.org/10.1093/ajcp/aqaa029.