Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Hospital Infection 107 (2021) 40–44

Available online at www.sciencedirect.com

Journal of Hospital Infection

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

Short Report

Clinical utility of SARS-CoV-2 whole genome sequencing in deciphering source of infection

T. Takenouchi [a,1], Y.W. Iwasaki [b,1], S. Harada [c], H. Ishizu [b], Y. Uwamino [d,e,f], S. Uno [d,e], A. Osada [a], K. Abe [g], N. Hasegawa [d,e], M. Murata [f], T. Takebayashi [c], K. Fukunaga [h], H. Saya [i], Y. Kitagawa [g], M. Amagai [j], H. Siomi [b,**], K. Kosaki [k,*], Keio Donner Project

[a] *Department of Pediatrics, Keio University School of Medicine, Tokyo, Japan*
[b] *Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan*
[c] *Department of Preventive Medicine and Public Health, Keio University School of Medicine, Tokyo, Japan*
[d] *Division of Infectious Diseases and Infection Control, Keio University Hospital, Tokyo, Japan*
[e] *Department of Infectious Diseases, Keio University School of Medicine, Tokyo, Japan*
[f] *Department of Laboratory Medicine, Keio University School of Medicine, Tokyo, Japan*
[g] *Department of Surgery, Keio University School of Medicine, Tokyo, Japan*
[h] *Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan*
[i] *Division of Gene Regulation, Institute for Advanced Medical Research, Keio University School of Medicine, Tokyo, Japan*
[j] *Department of Dermatology, Keio University School of Medicine, Tokyo, Japan*
[k] *Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan*

ARTICLE INFO

Article history:
Received 7 July 2020
Accepted 18 October 2020
Available online 24 October 2020

Keywords:
SARS-CoV-2
Whole genome sequencing
COVID-19
Nosocomial infection
Community infection

SUMMARY

Coronavirus disease 2019 (COVID-19) caused by human severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a worldwide problem. From the standpoint of hospital infection control, determining the source of infection is critical. We conducted the present study to evaluate the efficacy of using whole genome sequencing to determine the source of infection in hospitalized patients who do not have a clear infectious contact history. Recently, we encountered two seemingly separate COVID-19 clusters in a tertiary hospital. Whole viral genome sequencing distinguished the two clusters according to the viral haplotype. However, the source of infection was unclear in 14 patients with COVID-19 who were clinically unlinked to clusters 1 or 2. These patients, who had no clear history of infectious contact within the hospital ('undetermined source of infection'), had haplotypes similar to those in cluster 2 but did not have two of the mutations used to characterize cluster 2, suggesting that these 14 cases of 'undetermined source of infection' were not derived from cluster 2. Whole viral genome sequencing can be useful for

* Corresponding author. Address: Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo, 160-8582, Japan. Tel.: +81-3-5363-3754.
** Corresponding author. Address: Center for Medical Genetics, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo, 160-8582, Japan. Tel.: +81-3-5363-3890.
E-mail addresses: awa403@keio.jp (H. Siomi), kkosaki@keio.jp (K. Kosaki).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jhin.2020.10.014

confirming that sporadic COVID-19 cases with an undetermined source of infection are indeed not part of clusters at the institutional level.

Introduction

The current coronavirus disease 2019 (COVID-19) pandemic caused by human severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a major worldwide community problem. Since rapid increases in the numbers of COVID-19 patients threaten to collapse healthcare systems in many countries, the development of effective diagnostic and preventive systems is urgently needed.

SARS-CoV-2 is a single-strand RNA virus with a rapid pace of mutagenesis (approximately two new mutations per month) [1]. Currently, global monitoring of SARS-CoV-2 mutation dynamics is publicly available through the Global Initiative on Sharing All Influenza Data (GISAID) and Nextstrain [2,3]. Viral genome sequencing data is useful for tracing longitudinal and global trends in viral genome changes.

For infection control within hospitals, whole genome viral sequencing can help to determine whether newly diagnosed patients have nosocomial or community-acquired infections. In situations involving nosocomial infection, not only must confirmed positive cases be isolated, but thorough intrahospital contact tracing must be performed to identify healthcare workers and inpatients who may have undiagnosed COVID-19. In contrast, thorough intrahospital surveillance is not necessary in situations where infection is known to have occurred within the community. The roles of viral genomic data in the application of preventive measures at an institutional level remain to be explored.

Methods

Study population

The present study was conducted at Keio University Hospital, a single tertiary care medical centre in a metropolitan area (Tokyo, Japan). The present study protocol was approved by the ethics committee of the Keio University School of Medicine (approval number: 20200062). The hospital has a total of 960 beds with approximately 2700 workers, including 400 physicians. Among more than 80 university hospitals in Japan, Keio University Hospital was the first university hospital in Japan to be affected by a COVID-19 outbreak. Patients with a reverse transcription—polymerase chain reaction (RT—PCR)-positive result who had been diagnosed as having COVID-19 at Keio University Hospital between March 24th and May 15th, 2020, were enrolled in the present study. A total of 90 positive cases were identified (46 patients, 44 hospital staff). Among the 32 cases who underwent whole viral genome sequencing, 14 were patients and 18 were hospital staff. Cluster 1 occurred among the patients and hospital staff after one patient was transferred from a local hospital on March 19th, 2020. Cluster 2 occurred among interns and junior residents during the last week of March 2020. In addition, some cases were present that could not be linked to cluster 1 or 2 using contact tracing.

Clinical RT—PCR testing

A nasopharyngeal swab specimen was collected from individuals who were suspected of having COVID-19 based on the presence of fever, cough or rhinorrhoea, the appearance of pneumonia on computed tomography, or a history of close contact with a confirmed case. Clinical RT—PCR testing was performed using a standardized quantitative RT—PCR test for SARS-CoV-2 [4].

Specimen collection and sample preparation

The residual nasopharyngeal swab specimens of subjects who tested positive during clinical RT—PCR testing were retrospectively collected and used in the present analysis. Total RNA was extracted from the specimens using the QIAamp MinElute Virus Spin Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The RNA was reverse-transcribed to cDNA using a random hexamer primer and SuperScript III Reverse Transcriptase (Thermo Fisher, Waltham, MA, USA). PCR-based amplification was performed using Artic ncov-2019 primers, version 3, in two multiplex reactions according to the globally accepted 'nCoV-2019 sequencing protocol' [5,6]. A sequencing library for amplicon sequencing was prepared using the NEB Next Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA). Paired-end sequencing was performed on the MiSeq platform (Illumina, San Diego, CA, USA).

Bioinformatics analysis and annotation

The fastq files were aligned using the Burrows—Wheeler Aligner and a reference sequence (Wuhan-Hu-1, MN908947.3) to generate the bam files [7,8]. The bam files were then processed with iVar to remove primer positions supplied in a bed file and to soft clip primer sequences from an aligned and sorted bam file [9]. The quality of the genome sequencing data was evaluated using qualimap [10]. The sequenced bam files were processed with samtools and bcftools to call the variants in the variant call format (vcf) [11]. The variants were annotated for effects on protein translation and global viral allele frequencies using snpEff [2,12]. The bioinformatics pipeline used in this study, 'Variant calling pipeline for amplicon-based sequencing of the SARS-CoV-2 viral genome', is available at https://cmg.med.keio.ac.jp/sars-cov-2/.

Phylogenetic tree analysis

A phylogenetic tree analysis was performed locally using the Augur program available from Nextstrain and genome sequence data obtained in the currently reported study as well as data available from the global database EpiCov from GISAID [3]. For the construction of the phylogenetic tree, the analysis included all 32 subjects in the present study, the global dataset submitted as of February 29th, 2020, and all available Japanese

data excluding that obtained from the cruise ship, the *Diamond Princess* [13]. The allele frequency was calculated based on a total of 8604 sequences downloaded from the global database EpiCov (downloaded on April 16th, 2020).

Results

Ninety positive RT–PCR results were obtained during the study period at Keio University Hospital. Among these positive results, 32 corresponding samples were subjected to next-generation sequencing. The 'unrooted' analysis using Nextstrain showed that the presence of C11752T could be used to segregate the results into two distinctive clades. The first clade included five cases and was compatible with cluster 1. The second clade was divided into two groups by the presence of C823T and was compatible with cluster 2 and the 'undetermined source of infection' cases.

Cluster 1

The index patient was transferred from a local hospital to Keio University Hospital to undergo surgery on March 19th, 2020. The patient had no respiratory symptoms at admission. On March 23rd, 2020, the occurrence of a COVID-19 nosocomial infection at the local hospital prior to the patient's transfer was discovered. On the following day, the patient received a positive clinical RT–PCR test result [4]. Subsequently, three healthcare workers and four additional patients on the same floor tested positive using clinical RT–PCR testing.

Cluster 2

Another cluster occurred among interns and junior residents during the last week of March 2020. All 99 interns and junior residents underwent clinical RT–PCR testing, and 20 tested positive. The source of the infection remained unknown in this cluster.

Other patients unlinked to clusters 1 or 2

Concurrently with clusters 1 and 2, several patients who could not be linked to clusters 1 or 2 were newly diagnosed as having COVID-19. This group was considered to have an 'undetermined source of infection.'

Nasopharyngeal samples from five subjects in cluster 1, 13 subjects with positive RT–PCR results in cluster 2, and 14 subjects with an 'undetermined source of infection' were subjected to whole viral genome sequencing.

Phylogenetic tree analysis

The data points from cluster 1 were distinct from the data points from cluster 2 (Figure 1). Cluster 1 appeared to have derived from the original SARS-CoV-2 descended from the Wuhan outbreak at a relatively early stage, whereas the data points from the other Japanese COVID-19 cases, including those in cluster 2 and the 'undetermined source of infection' group, were clustered together rather closely.

The viral genome haplotype analysis confirmed that cluster 1 and cluster 2 were distinguished by 15 mutations (Figure 2).

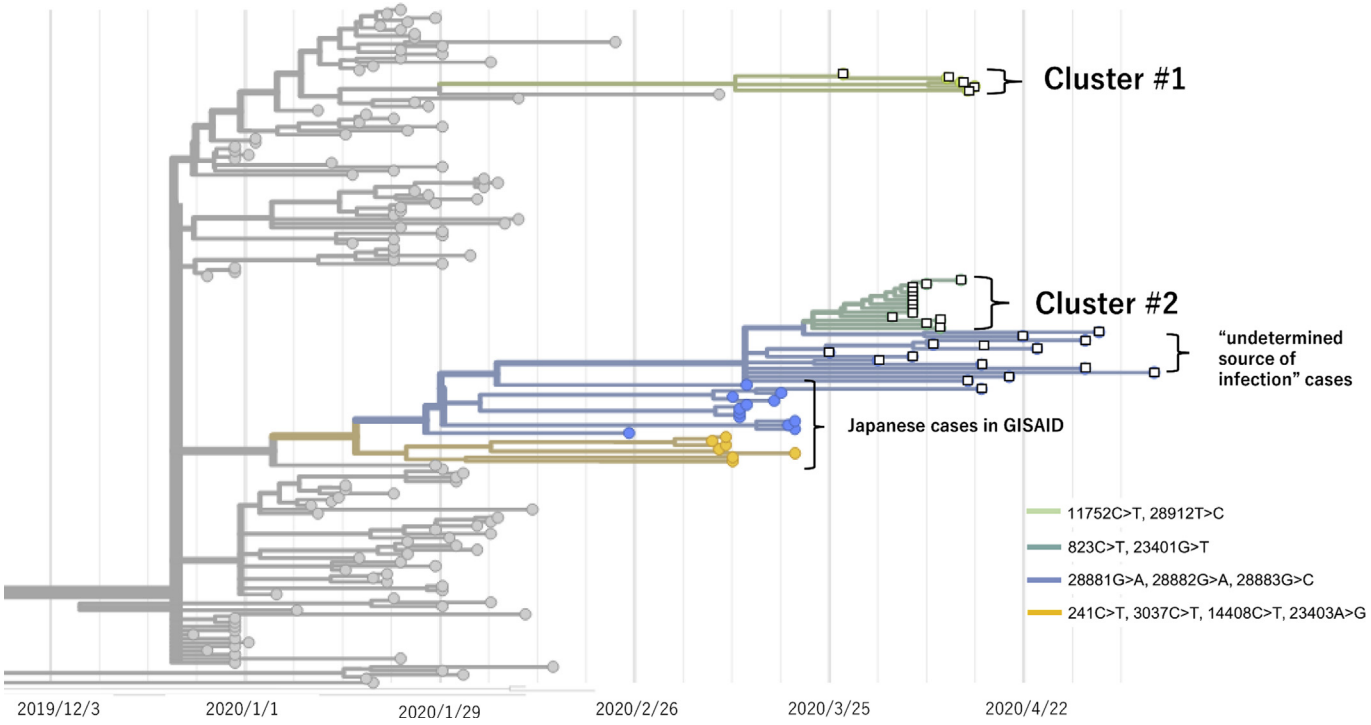

Figure 1. Results of phylogenetic tree analysis. The dots represent publicly available data points. The squares represent cases in the present study. Clades were defined according to the colour code shown at the bottom right. Note that the data points from cluster 1 and from cluster 2 were distinct. Cluster 2 and the 'undetermined source of infection' cases, as well as the Japanese cases registered in GISAID, formed clusters belonging to a close branch (see Figure 2).

The mutations 11752C→T, 25665C→T, 26447C→T, 27700-27702delATT, and 28912T→C were specific to cluster 1, whereas 241C→T, 313C→T, 823C→T, 3037C→T, 14408C→T, 23401G→T, 23403A→G, 28881G→A, 28882G→A, and 28883G→C were specific to cluster 2. The mutually exclusive haplotypes of clusters 1 and 2 provided molecular evidence that clusters 1 and 2 were caused by two different SARS-CoV-2 strains, and thus were independent of each other. This finding was compatible with the in-hospital surveillance results obtained using contact tracing.

Viral genome haplotype of 'undetermined source of infection' cases

During the same period as the two in-hospital clusters, 14 cases with an 'undetermined source of infection' occurred, necessitating an urgent determination of the source of infection. Although these 14 subjects had a haplotype that was similar to that of cluster 2, a distinctive viral genomic signature was also present: two mutations, i.e. 823C→T and 23401G→T, were specific to the cases in cluster 2, but not to the 'undetermined source of infection' cases. This finding suggested that the 'undetermined source of infection' cases had community-acquired infections and were not derived from cluster 2, since the spontaneous reversion of the viral genome mutations was unlikely. Instead, the cluster 2 and 'undetermined source of infection' cases were most likely derived from a common ancestral haplotype with eight mutations, i.e. 241C→T, 313C→T, 3037C→T, 14408C→T, 23403A→G, 28881G→A, 28882G→A, and 28883G→C (Figure 2), present in the neighbourhood surrounding Keio University Hospital.

The utility of viral haplotype analysis is best exemplified by the case of one healthcare worker (case 23, Supplementary Table S1) in the 'undetermined source of infection' group. The subject worked as a full-time employee at Keio University Hospital. In addition, until March 17th, 2020, she had attended an outpatient clinic once a week at the local hospital that was the origin of cluster 1. She developed a fever and tested positive using clinical RT–PCR on April 13th, 2020. She denied having had any contact with the individuals from clusters 1 or 2. Her viral genome haplotype confirmed that she had not become infected with SARS-CoV-2 at either the local hospital or Keio University Hospital.

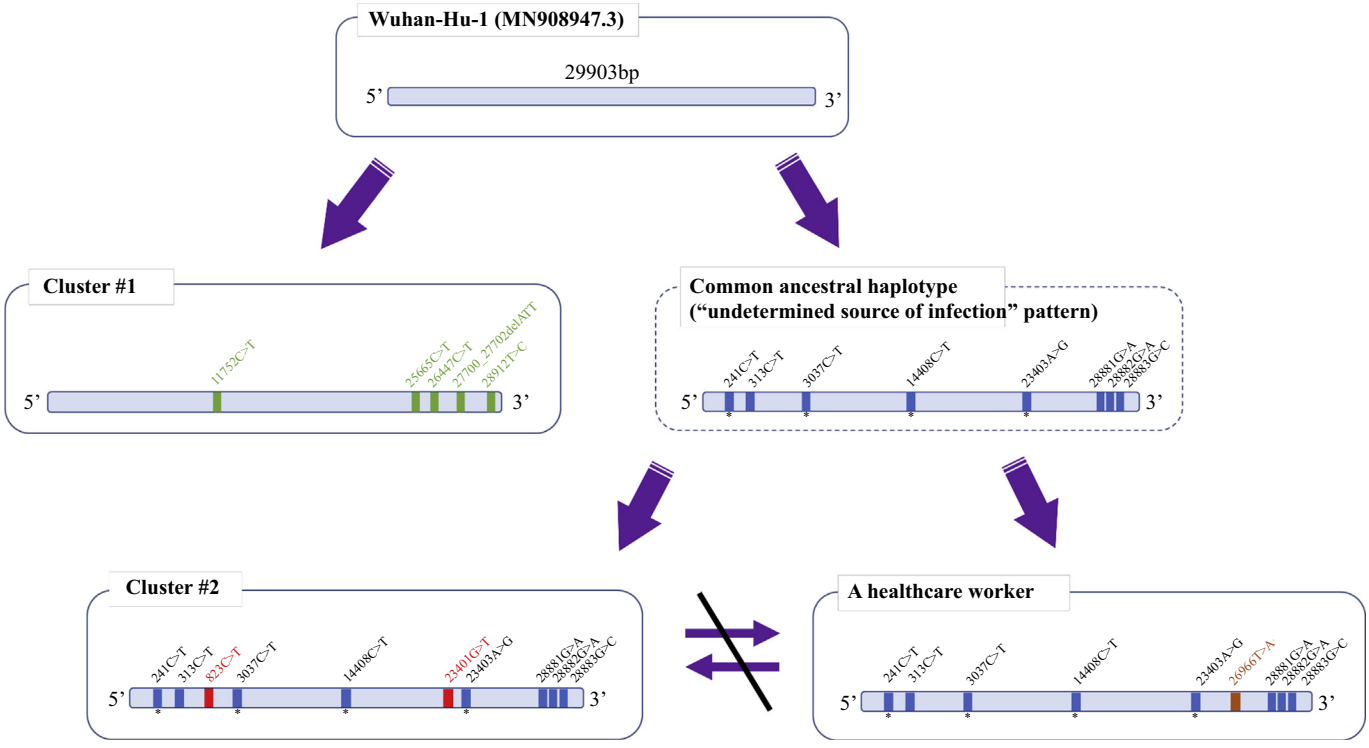

Figure 2. Postulated local evolutionary history based on viral genome haplotypes. The asterisks indicate common mutations that were present in approximately 56% of publicly available 8604 SARS-CoV-2 cases downloaded from the GISAID database as of April 16th, 2020. Note that five mutations, i.e. 11752C→T, 25665C→T, 26447C→T, 27700_27702delATT, and 28912T→C, were exclusively present in cluster 1 and were absent in cluster 2. Conversely, ten mutations, i.e. 241C→T, 313C→T, 823C→T, 3037C→T, 14408C→T, 23401G→T, 23403A→G, 28881G→A, 28882G→A, and 28883G→C, were exclusively present in all 13 subjects from cluster 2, whereas none of these changes were present in subjects from cluster 1. The haplotypes of cluster 2 and all the cases in the 'undetermined source of infection' group shared eight mutations, i.e. 241C→T, 313C→T, 3037C→T, 14408C→T, 23403A→G, 28881G→A, 28882G→A, and 28883G→C; two mutations, 823C→T and 23401G→T, were not shared. The representative case of a healthcare worker, shown at the bottom right, did not exhibit either 823C→T or 23401G→T but contained 26966T→A, which was not present in either cluster 1 or cluster 2. This observation indicated that the strains carried by this healthcare worker and those in cluster 2 were likely derived from a shared ancestor, i.e. 'a community infection pattern,' rather than direct cross-infection with each other.

Discussion

In the present study, rapid onsite whole viral genome sequencing of SARS-CoV-2 successfully demonstrated distinctive viral genomic haplotypes that were concordant with the epidemiologic contact history in two intrahospital clusters. The viral genome haplotype was useful for confirming that sporadic COVID-19 cases with an undetermined source of infection were indeed not part of clusters within the hospital environment. Epidemiologic contact tracing combined with viral genomic data could be effective as a preventive measure against COVID-19.

The major limitations of the present study were the relatively small number of subjects, the inclusion of a single medical centre, and the lack of a systematic method of subject accrual. Since the virus—host interaction, which determines virulence and the severity of symptoms, is highly complex and multifactorial, co-analyses of host genome data are needed to determine the relationship between viral genomic data and prognosis and therapeutic efficacy.

The acquisition of whole viral genome sequences has implications from a future basic research perspective. In the present study, the whole viral genome sequencing not only provided the nucleotide signatures of the SARS-CoV-2 strains, but also identified 53 different mutations, with 27 being amino acid substitutions, in 32 samples. A recent longitudinal observation of the SARS-CoV-2 genome has shown that a single mutation in the spike region, i.e. D614G, became predominant in early 2020 and increased the amount of viral nucleic acid shedding [14]. The present study showed that cluster 1 exhibited D614, whereas cluster 2 and the 'undetermined source of infection' cases exhibited G614; these findings were compatible with the present epidemiological observations. The research and development of vaccines and antibodies targeting SARS-CoV-2 should be pursued in view of this variability in viral protein sequences.

In conclusion, we have shown that whole viral genome sequencing was useful for confirming that sporadic COVID-19 cases with an undetermined source of infection were indeed not part of institutional clusters. Whole viral genome sequencing is useful for augmenting the results of thorough contact tracing, particularly in situations where multiple COVID-19 clusters have occurred within a single hospital simultaneously.

Acknowledegements

We downloaded the full nucleotide sequences of the SARS-CoV-2 genomes from the GISAID database (https://www.gisaid.org/). We uploaded the full nucleotide sequences of our cohort to the GISAID database. A table of the contributors is available in Supplementary Table S1. We thank all the patients and healthcare workers who have fought against COVID-19. The Keio Donner Project is devoted to the late Professor S. Kitasato, the founder of The Keio University School of Medicine.

Conflict of interest statement
None declared.

Funding sources
This work was supported by Keio Gijuku Academic Development Funds and by AMED under Grant Number JP20he0622043.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhin.2020.10.014.

References

[1] Penarrubia L, Ruiz M, Porco R, Rao SN, Juanola-Falgarona M, Manissero D, et al. Multiple assays in a real-time RT—PCR SARS-CoV-2 panel can mitigate the risk of loss of sensitivity by new genomic variants during the COVID-19 outbreak. Int J Infect Dis 2020;97:225—9.
[2] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 2017;1:33—46.
[3] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121—3.
[4] Shirato K, Nao N, Katano H, Takayama I, Saito S, Kato F, et al. Development of genetic diagnostic methods for novel coronavirus 2019 (nCoV-2019) in Japan. Jpn J Infect Dis 2020;73:304—7.
[5] Artic ncov-2019 primers, version 3. 2020. https://artic.network/resources/ncov/ncov-amplicon-v3.pdf
[6] Quick J. nCoV-2019 sequencing protocol, vol. 1; 2020. https://dx.doi.org/10.17504/protocols.io.bbmuik6w.
[7] Li H, Durbin R. Fast and accurate short read alignment with Burrows—Wheeler transform. Bioinformatics 2009;25:1754—60.
[8] 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.
[9] 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.
[10] Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 2016;32:292—4.
[11] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078—9.
[12] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80—92.
[13] Moriarty LF, Plucinski MM, Marston BJ, Kurbatova EV, Knust B, Murray EL, et al. Public health responses to COVID-19 outbreaks on cruise ships — worldwide, February—March 2020. Morb Mortal Wkly Rep 2020;69:347—52.
[14] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182:812—827 e19.